IMPERIAL

Imperial College London

Department of Mathematics

Change Point Detection

Author: Hugo Ng (CID: 01848839)

A thesis submitted for the degree of

MSc in Mathematics and Finance, 2023-2024

Declaration

The work contained in this thesis is my own work unless otherwise stated.

Acknowledgements

First of all, I would like to thank my supervisor, Dr. Cris Salvi, for his invaluable guidance and support throughout this project. I'm also deeply grateful to Aitor Muguruza, Zan Zuric, and James Mc Greevy from Kaiju Capital Management for their insights and the opportunity to apply my research in a practical setting.

Finally, I want to extend my gratitude to my family and friends for their unwavering support and encouragement.

Abstract

In this paper, we develop change point detection methods for financial time series, with a focus on non-parametric approaches that provide additional flexibility compared to traditional methods. We empirically evaluate the performance of the proposed algorithms and demonstrate their effectiveness on both synthetic and real-world data. Our work builds upon non-parametric two-sample tests, particularly the maximum mean discrepancy (MMD) statistic, which has shown robust performance in detecting distributional shifts. We extend this concept to develop an ϵ-real time, rolling window approach for detecting market regimes when the number of change points is unknown. Finally, we explore the potential of using detected change points as labels for predictive modelling, aiming to forecast future regime shifts. We further discuss the implications of such methods in risk management and portfolio optimisation.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The detection of regime changes in financial markets is a critical task for risk management, portfolio optimization, and trading strategy development. Ang and Timmermann [2] provided a comprehensive overview of regime changes in financial markets, discussing their implications for asset pricing, portfolio choice, and risk management. They emphasized how regime shifts in fundamental processes (e.g., consumption or dividend growth) significantly affect equilibrium asset prices and risk-return trade-offs. Kritzman, Page, and Turkington [26] presented practical implementations of markov-switching models for dynamic investment strategies. They demonstrated how using dynamically adjust portfolio allocation outperformed static allocations, especially in avoiding large losses during turbulent periods. These studies highlight the importance of developing regime detection methods to develop reactive trading strategies. Traditionally, this problem has been addressed using Hidden Markov Models (HMMs), which rely on assumptions about latent state variables and parametric distributions [19]. However, model-dependent approaches may not capture the full complexity of financial time series, especially during periods of market turbulence or structural shifts.

Recent advancements in the market regime clustering problem utilise unsupervised learning techniques and have shown promise in addressing these limitations. Notably, k-means clustering with p-Wasserstein distance [24] has demonstrated superior performance compared to moment-based k-means and HMMs, when applied to real data with known periods of instabilities and on synthetic data. The idea has been extended to multivariate market regimes [30], taking into account the cross-correlation between assets when forming clusters. These developments sparked our interest in developing a more flexible, non-parametric approach to market regime detection, where the number of change points is unknown.

While existing libraries like Ruptures [36] provide tools for offline change point detection using optimization approaches to balance a cost function that quantifies differences between segments with a penalty function penalising the complexity of the model[37, 32], our work focuses on developing a non-parametric method based on two-sample tests [25, 14]. This approach offers several advantages, including model-free analysis and statistical robustness of the identified regimes.

Non-parametric change point detection methods can be broadly categorized into likelihood-based [40], that relies on maximising the distribution-free log likelihood, rank-based[28], and kernel-based approaches [20, 21, 7]. In this work, we focus on exploring kernel-based techniques, particularly maximum mean discrepancy (MMD) two-sample tests [17], due to their demonstrated power and growing popularity in detecting distributional changes without imposing restrictive assumptions on the underlying data-generating process.

In this paper, we contextualize the market regime detection problem within the broader framework of change point detection – the identification of abrupt shifts in the probability

distribution of a stochastic process or time series. We first provide background information and establish the theoretical framework for our approach. Then, we consider offline detection scenarios, where the entire dataset is available for analysis, before proceeding to online detection, which processes data sequentially in real-time. Additionally, we explore the potential of using detected change points as labels for predictive modeling, aiming to forecast future regime shifts.

1.1 Background

The foundation of our analysis in this report is built upon two-sample tests, as they play a crucial role in our framework of change point detection. A change point differentiates two segments of data exhibiting distinct statistical properties. Two-sample tests offer a rigorous statistical framework for identifying regime shifts or other significant changes in financial time series. By comparing the distributions of two subsegments, it enables us to quantify the likelihood of shifts in market dynamics. Hence, we introduce the following definition of two-sample test.

Definition 1.1.1. (Two-sample test, [34], Section 2). Let ($\mathcal{X}$, d) be a metric space, and p, q be two Borel probability measures defined on $\mathcal{X}$. Given random variables $x \overset{i.i.d.}{\sim} p$ and $y \overset{i.i.d.}{\sim} q$ and $X = (X_i)_{i=0}^n$, $Y = (Y_i)_{i=0}^m$, the corresponding realisations, a two-sample test $\delta(X, Y) : \mathcal{X}^n \times \mathcal{X}^m \to \{0, 1\}$ is used to distinguish between

$$H_0 : p = q \quad \text{against} \quad H_1 : p \neq q$$

The test function δ can be further described by an indicator function which returns 1 when the test statistic exceeds a predetermined threshold and 0 otherwise. The test is said to be level α when the type I error $P_{H_0}(\delta(X, Y) = 1) \leq \alpha$. The inequality can be inverted to obtain the rejection threshold. The power of the test is defined to be $P_{H_1}(\delta(X, Y) = 1)$ i.e. the probability that the null hypothesis is correctly rejected. A test is consistent if the power increases to 1 in the limit of both sample size.

Other common two sample tests include the Kolmogorov-Smirnov test and Cramer-von Mises test which both draw test statistics based on the distance between two empirical cumulative density functions of the samples. While they are non-parametric, they generally suffer from the curse of dimensionality and the adaptation to multivariate version is often challenging, see [27]. More recently, Gretton et al. [17] proposed a kernel two-sample test based on MMD test statistics over a unit ball in the reproducing kernel Hilbert space (RKHS). And it has been shown that the kernel two-sample test consistently outperforms a range of parametric and non-parametric tests in a multidimensional setting. We first provide the definition of a reproducing kernel Hilbert space and a reproducing kernel.

Definition 1.1.2. (Reproducing kernel Hilbert space, [38]). Let ($\mathcal{X}$, d) be a separable measurable metric space. A Hilbert space ($\mathcal{H}$, $\langle \cdot, \cdot \rangle_{\mathcal{H}}$) is a reproducing kernel Hilbert space if $\forall x \in \mathcal{X}$, the point evaluation operator δ_x which maps $f \in \mathcal{H}$ to $f(x) \in \mathbb{R}$ is bounded i.e. $\exists M_x$ such that

$$|\delta_x(f)| \leq M_x \|f\|_{\mathcal{H}} \qquad \forall f \in \mathcal{H}.$$

By Riesz representation theorem, for all $x \in \mathcal{X}$ there exists an element $\phi(x) \in \mathcal{X}$ such that

$$\delta_x(f) = f(x) = \langle \phi(x), f \rangle_{\mathcal{H}} \quad \forall f \in \mathcal{H}.$$

Definition 1.1.3. (Reproducing kernel, [38]) A reproducing kernel $k(x, y) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ of $\mathcal{H}$ is defined by

$$k(x, y) := \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}.$$

The reproducing kernel defined in Definition 1.1.2 is clearly symmetric and positive definite. More importantly, the converse statement is also true.

Theorem 1.1.4. *(Moore-Aronszajn theorem, [3]). If $k(x, y) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a symmetric positive definite kernel, then there exists a unique Hilbert space on $\mathcal{X}$ such that k is a reproducing kernel.*

This allows any kernel k satisfying the above properties to define its unique corresponding RKHS.

1.1.1 Maximum Mean Discrepancy

Lemma 1.1.5. *([10]). With notations same as in Definition 1.1.1, $p = q$ if and only if $\mathbb{E}_p[f(x)] = \mathbb{E}_q[f(y)]$ for all $f \in \mathcal{C}$ where $\mathcal{C}$ is the space of bounded continuous functions.*

Motivated by Lemma 1.1.5, the following statistic was proposed to quantify the disparity between measures p and q.

Definition 1.1.6. (Maximum mean discrepancy, [17], Definition 2). Let $\mathcal{F}$ be a class of functions $f : \mathcal{X} \to \mathbb{R}$. The maximum mean discrepancy is defined as

$$\mathrm{MMD}[\mathcal{F}, p, q] := \sup_{f \in \mathcal{F}} (\mathbb{E}_p[f(x)] - \mathbb{E}_q[f(y)])$$

Notice that the MMD statistic is the supremum taken over the function class $\mathcal{F}$. Although there are various options for the choice of function class, the RKHS $\mathcal{H}$ is frequently used due to its rich function space, uniquely identifying whether p = q with finite samples. Additionally, the use of kernel allows for efficient computation of inner products in high-dimensional spaces without explicitly mapping the data to those spaces, known as the "kernel trick". The range of reproducing kernels also enables the RKHS to be tailored to different types of data and distributions, providing flexibility for solving a wide range of problems.

Definition 1.1.7. (Mean embedding of measure p, [12], Section 2.2). The mean embedding μ_p of measure p is a unique element in RKHS $\mathcal{H}$ such that $\mathbb{E}_p f = \langle f, \mu_p \rangle$ for all $f \in \mathcal{H}$.

Lemma 1.1.8. *([17], Lemma 4). If the reproducing kernel $k(\cdot, \cdot)$ of RKHS $\mathcal{H}$ is measurable and $\mathbb{E}_p[\sqrt{k(x, x)}] \leq \infty$, the mean embedding of p exists. Furthermore,*

$$\mathrm{MMD}^2[\mathcal{H}, p, q] = \|\mu_p - \mu_q\|^2_{\mathcal{H}}.$$

Proof in Appendix A.

Proposition 1.1.9. *Let x' and y' be independent copies of x and y respectively and n, m the corresponding sample sizes, the squared population MMD can be expressed as*

$$\mathrm{MMD}^2[\mathcal{H}, p, q] = \mathbb{E}_{x,x'}\left[k\left(x, x'\right)\right] - 2\mathbb{E}_{x,y}[k(x, y)] + \mathbb{E}_{y,y'}\left[k\left(y, y'\right)\right].$$

Hence the unbiased empirical estimate is given by,

$$\begin{aligned}\mathrm{MMD}^2_u[\mathcal{H}, X, Y] =& \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} k\left(X_i, X_j\right) + \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j \neq i}^{m} k\left(Y_i, Y_j\right) \\ & - \frac{2}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} k\left(X_i, Y_j\right).\end{aligned} \tag{1.1.1}$$

Proof.

$$\begin{aligned}\mathrm{MMD}^2[\mathcal{H}, p, q] &= \|\mu_p - \mu_q\|^2_{\mathcal{H}} \\ &= \langle \mu_p, \mu_p \rangle_{\mathcal{H}} + \langle \mu_q, \mu_q \rangle_{\mathcal{H}} - 2 \langle \mu_p, \mu_q \rangle_{\mathcal{H}} \\ &= \mathbb{E}_{x,x'} \left\langle \phi(x), \phi\left(x'\right) \right\rangle_{\mathcal{H}} + \mathbb{E}_{y,y'} \left\langle \phi(y), \phi\left(y'\right) \right\rangle_{\mathcal{H}} - 2\mathbb{E}_{x,y} \langle \phi(x), \phi(y) \rangle_{\mathcal{H}},\end{aligned}$$

The first equality follows from Lemma 1.1.8. Since $\langle \phi(x), \phi(x') \rangle_{\mathcal{H}} = k(x, x')$, the result follows. Lastly, the empirical estimate is obtained by replacing the expectations with unbiased sample averages estimators. □

Since we would like the MMD to be zero when the measures p and q are identical, it is important for the MMD to be a metric. It has been shown that MMD is indeed a metric if the associated kernel is universal on a compact space $\mathcal{X}$. Otherwise, in the case where $\mathcal{X}$ is not compact, MMD remains a metric if the kernel is characteristic, see Appendix A.

Definition 1.1.10. (Characteristic kernel, [13], Section 2.2). Let $\mathcal{P}$ be the family of all probability measure on ($\mathcal{X}$, d), kernel k is called characteristic if the mapping

$$\mathcal{M}_k : \mathcal{P} \to \mathcal{H}$$

i.e. p is mapped to μ_p, is injective. Hence, $\mathbb{E}_p[f(x)] = \mathbb{E}_q[f(y)]$ $\forall f \in \mathcal{H}$ implies $p = q$. Analogous to the characteristic function, this defines a kernel with mean embeddings that uniquely determine probability measures on $\mathcal{X}$.

The two commonly used characteristic kernels are the Gaussian kernel

$$\kappa_G = \exp(-\frac{1}{2\sigma^2}\|x - y\|^2)$$

and the Laplacian kernel

$$\kappa_L = \exp(-\frac{1}{\sigma^2}\|x - y\|).$$

And we will compare their performances in later sections.

Chapter 2

Offline Detection

2.1 Kernel two sample test

This section reviews the work of Gretton et al. and provides related experimental results. The first hypothesis test based on unbiased statistics follows from Theorem 2.1.1. The U statistics deviation bound [23], is applied in the context of MMD statistics and yields the following.

Theorem 2.1.1. *(Bound for unbiased MMD statistics, [17], Theorem 10). Assume $k(x_i, x_j)$ is bounded above by K,*

$$P(\mathrm{MMD}_u^2[\mathcal{F}, X, Y] - \mathrm{MMD}^2[\mathcal{F}, p, q] > t) \leq \exp\left(\frac{-t^2 m_2}{8K^2}\right)$$

where $m_2 := \lfloor m/2 \rfloor$ and m is the sample size of X, Y.

Under the null hypothesis, $p = q$, i.e. $\mathrm{MMD}^2[\mathcal{F}, p, q] = 0$, the acceptance region of test with level α is given by

$$\mathrm{MMD}_u^2[\mathcal{F}, X, Y] < \frac{4K}{\sqrt{m}}\sqrt{\log \alpha^{-1}}.$$

However, it is pointed out in [17] that this bound is conservative by design as it does not take into account the actual distribution p, q and has a finite sample size. Also, it has been shown empirically when distribution p, q are Gaussian with different mean or variance, this bound yields much worse performance compared to the t-test, Kolmogorov-Smirnov test and other variants of the MMD test.

In practice, the asymptotic distribution of the unbiased test statistics is often considered. As the unbiased estimate of MMD converges asymptotically to a weighted infinite series of χ^2 distributions under the null hypothesis, approximations of the $1 - \alpha$ quantile falls into one of two approaches. The first method employs bootstrapping or permutation techniques. By resampling or shuffling the original samples x_i and y_i, we generate new sets of samples that are statistically drawn from the same distribution. The MMD estimates computed from these bootstrapped samples converge consistently to the true null distribution. The second method involves fitting known distributions to the moments of the null distribution. By matching the moments of the empirical null distribution to a family of distributions, such as the Pearson curve, we can obtain the theoretical quantiles for hypothesis testing.

2.1.1 Comparison of power

Here we present a comprehensive comparison of various two-sample tests through experimental results. We evaluate the power of these tests using a Monte Carlo approach,

defined as the proportion of correct rejections under the alternative hypothesis over 100 repetitions. Our experiments cover a range of alternative scenarios, including changes in the first four moments, to assess the sensitivity of each tests in detecting different types of distributional shifts. For the MMD tests, unless otherwise stated, we estimated the $1 - \alpha$ quantile using 1000 iterations of permutation resampling. Also, throughout this analysis, we maintained a significance level of $\alpha = 0.05$.

- Mean Shift

  When samples were drawn from two normal distributions with different means, the Student's t test (t) [39] demonstrated superior power across all sample sizes. This aligns with our expectation, given that the t test is a parametric test specifically designed to identify changes in mean for normally distributed data. However, when applied to comparing gamma distributions with identical variances but different means, the performance deteriorated significantly. Results are provided in Appendix B. As shown in Figure 2.1, all tested methods produced similar performance and exhibited consistency. When the sample size increased, the statistical power approached 1. This consistency allows us to draw confident conclusions and reliably distinguish true distributional changes when sufficient data is given. To ensure that the tests are valid and well-defined, we estimated the Type I error rates following similar procedure to the power tests, with the key difference being that both sample sets were drawn from N(1, 1), representing the case of no distribution change. As all methods maintained error rates close to the 5% level, we confirmed that they are all well defined level α tests. Detailed numerical results for Type I errors are provided in Appendix B.

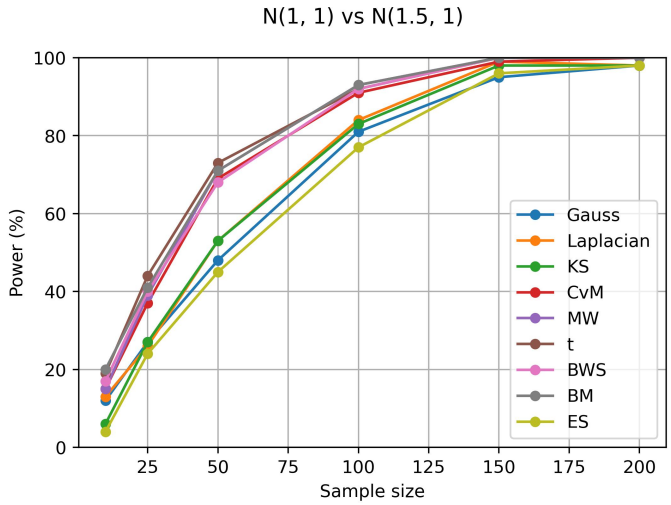


Figure 2.1: Power (%) against sample size at level $\alpha = 5\%$

- Variance Change

  For detecting changes in variance between two normal distributions with equal means, the performances of the tests varied significantly. The Epps-Singleton (ES) test [16, 11] demonstrated the highest power across almost all sample sizes, closely followed by the Gauss and Laplacian tests. The Baumgartner-Weiss-Schindler (BWS) test [31] also showed strong performance, especially as sample size increased. The Kolmogorov-Smirnov (KS) [27] and Cramer-von Mises (CvM) tests [1] had moderate power that improved with larger sample sizes. Notably, the Student's t-test, Mann-Whitney (MW) [29], and Brunner-Munzel (BM) tests [5] showed very low power

regardless of sample size. The unsatisfactory performance of t-test is expected as it is primarily designed to detect changes in location rather than scale. All other tests exhibited increasing power with larger sample sizes, demonstrating consistency. The ES, Gauss, and Laplacian tests achieved near-perfect power (close to 100%) with relatively small sample sizes (around 50-75), indicating their efficiency in detecting variance changes in normal distributions.

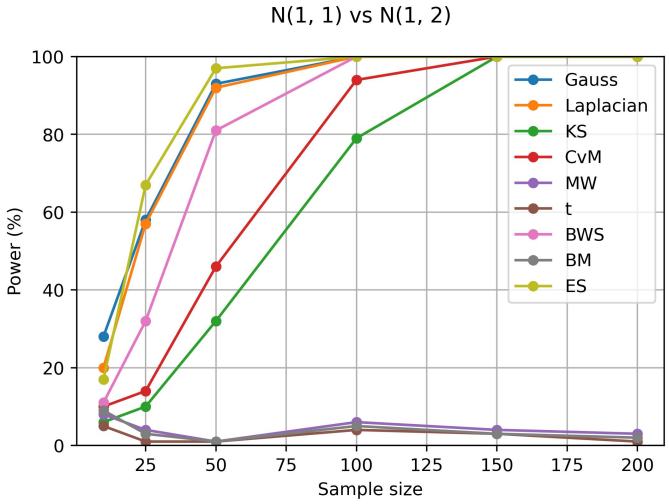


Figure 2.2: Power (%) against sample size at level $\alpha = 5\%$

- Kurtosis Change

  When comparing standard normal against standardised Student's t-distribution with 2 degrees of freedom, we observed a wide range of performances in Figure 2.3. While the Epps-Singleton test demonstrated near perfect accuracy under the alternative hypothesis, followed closely by Gauss and Laplacian MMD test, the Student's t-test, Mann Whitney (MW) and Brunner-Munzel(BM) again performed worst across all sample sizes. The Cramer-von Mises (CvM) and Kolmogorov-Smirnov (KS) tests displayed moderate power that improved gradually with increasing sample size. As it is common to observe deviations from normality to heavier tailed distributions in actual log return, the effectiveness in detecting kurtosis or higher moment differences is crucial in making sound statistical inferences on financial time series change points.

- Skewness Change

  To test for changes in skewness, we used the skew normal distribution, parameterized by location ξ, scale ω, and shape α. We carefully selected ξ and ω to ensure that the resulting distribution had a mean of 0 and variance of 1. With a shape parameter $\alpha = 4$, the distribution exhibited a skewness of 0.78 and an excess kurtosis of 0.63. Our results, as illustrated in Figure 2.4, revealed that the Epps-Singleton test and the MMD tests with Gaussian and Laplacian kernels significantly outperformed other methods in detecting skewness changes. However, changes in skewness were shown to be the most challenging to detect among all scenarios tested. Even the best-performing tests required nearly 500 samples to yield high confidence in determining a change. It is also worth noting that for the 1000 sample size scenario, we used 100 permutations to bootstrap the null distribution in order to maintain reasonable computation time.

- Bullish to Bearish Shift

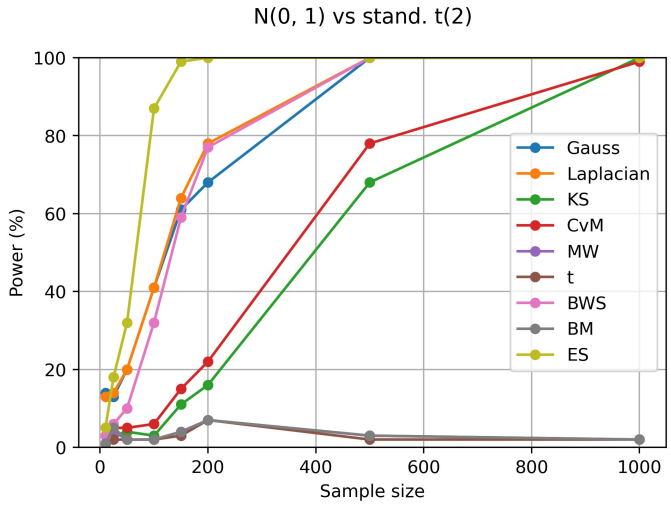


Figure 2.3: Power (%) against sample size at level $\alpha = 5\%$

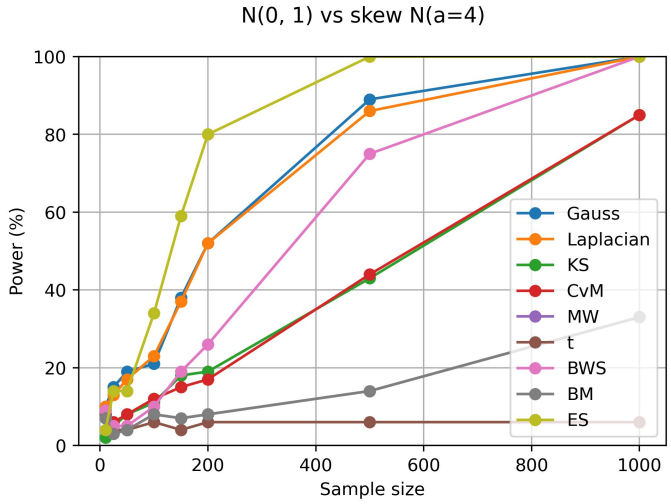


Figure 2.4: Power (%) against sample size at level $\alpha = 5\%$

To ensure robustness when applying two-sample tests in financial time series, we wish to investigate the accuracy of the above tests under realistic distributions. Theodoussiou et al. [35] proposed a family of distributions called Skewed Generalised t (SGT) that captures a range of commonly used distributions in statistics including normal, Student's t, Cauchy and uniform distribution. It was originally proposed to fit stock returns. In this experiment, we used parameters similar to the fit of S&P 500 and Topix in Table 6 of [35] to represent the bull and bear market conditions respectively as they coincided with usual observation that bull markets have positive mean with smaller variance while bear markets have negative mean with larger variance. Exact values were chosen as follows.

$$\theta_{\text{bull}} = (0.7, 0.02, 1.5, 9)$$

and

$$\theta_{\text{bear}} = (1.2, -0.02, 1.4, 5).$$

Results in Figure 2.5 revealed similar disparities in test performance compared to the ones in moment changes we presented. The power of tests displayed almost linear increase as the sample size growed, with the exception of the Student's t,

Mann Whitney (MW) and Brunner-Munzel(BM) tests. They had a rejection rate less than 10% under the alternative hypothesis. The Epps-Singleton (ES) and MMD tests again outperformed the other tests with power exceeding 90% at sample size of 200.

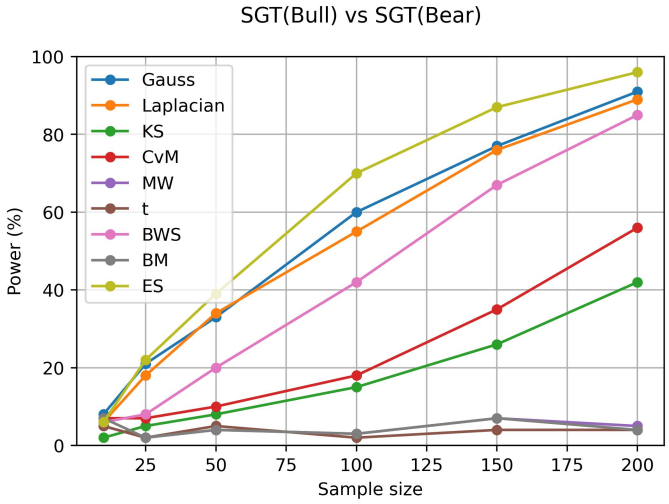


Figure 2.5: Power (%) against sample size at level $\alpha = 5\%$

The series of power comparison tests conducted across various distributional scenarios demonstrated several key findings. We consistently observed that the power of all tests increased with sample size, highlighting the importance of adequate data for reliable detection of distributional changes. We also verified that the MMD two-sample tests, along with other established tests, are well-defined, meaning that the empirical type I error rate matched the expected 5% under the null hypothesis. Notably, the Epps-Singleton test and MMD tests with Gaussian and Laplacian kernels consistently outperformed other methods across all scenarios. The superior performance was especially evident when comparing distributions that differed in shape rather than location or scale, such as the cases of heavy-tailed or skewed distributions. These experimental results provide crucial insights into the relationship between sample size and statistical confidence, offering a framework to determine the appropriate window size when applied to real-world data with unknown underlying distributions.

2.1.2 Effect of sample size ratio

In addition to the above study, we extend our investigation to the impact of sample size ratio on the power of two sample tests. This is particularly relevant for practical applications, especially in financial data. In real-world settings, we often have access to a larger amount of historical data compared to future data when attempting to identify regime shifts. Having an imbalanced sample ratio might be beneficial as increasing the future window length can lead to longer detection delays, compromising the usefulness of the identified change points.

To examine this effect, we performed a series of tests with varying sample size ratios between the two groups being compared. In Figure 2.6, n_1 refers to the number of samples in the first data set while the number of samples in the second dataset n_2 is given by n_1 divided by the corresponding ratio.

Several key observations can be drawn from the results in Figure 2.6. Firstly, increasing the sample size ratio for fixed n_1 generally led to reduced test power across all base sample sizes n_1. As the total number of samples decreased, it is reasonable to observe reduction

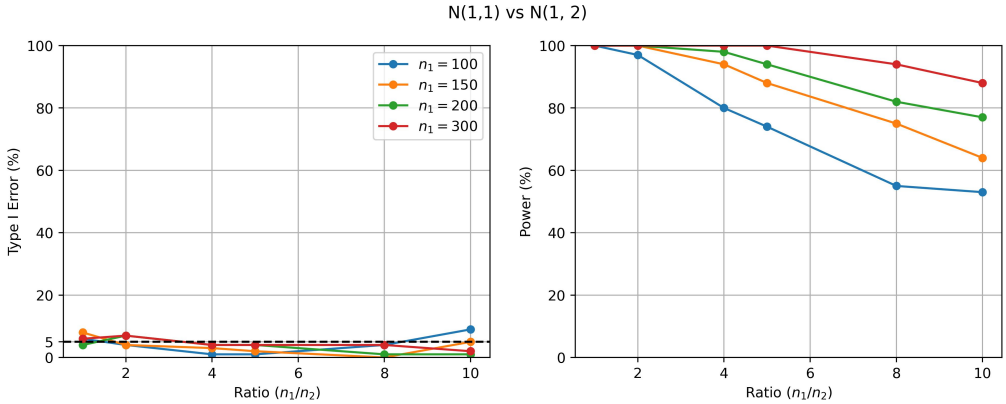


Figure 2.6: Type I error and power (%) against ratio between sample sizes of two sets of independent observations at level $\alpha = 5\%$

in performance. Also, we noticed that the impact of increasing sample ratio was minimal when n_1 is large, even a tenfold decrease in the smaller sample set yielded almost identical performance. In contrast, when n_1 was relatively small, a decrease by fourfold already resulted in significant decline in power.

These findings highlighted the fact that the power of two sample test is both a function of the sample ratio and total sample size. Yet, if we apply this finding from a different perspective, one can conclude that when the number of samples in the smaller set, usually being the future window, is small, increasing the number of samples in the past and larger window will increase the power of the test. Thus, it provides a valuable insight to practical application.

It is worth noting that the above figure illustrates the effect for changes in variance. We observed similar patterns for mean changes as well, and the corresponding plot can be found in Appendix B.

2.1.3 Optimal bandwidth

In this section, we will comment on the effect of the choice of kernel bandwidth to the power and accuracy of the two sample test. In general, the performance of kernel-based methods, including the MMD test, is heavily dependent on the choice of bandwidth σ. A commonly used approach for bandwidth selection is the median heuristic, which is the median of the pairwise distances in the pooled sample. As described in [15], the rationale behind this approach is to set the parameter σ in the same order as the entries of the Gram matrix K. At the two extremes $\sigma \to 0, \infty$, the MMD statistic becomes 0, losing all relevant information. Hence, a simple approach is to choose σ as the middle range of all pairwise distances.

While the median heuristic offers a simple and efficient solution, it is important to note that it is not theoretically optimal for maximising the test power. We will illustrate this idea with the following experiment.

In Figure 2.7 , we provide the power of MMD two-sample tests using Gaussian kernel with different multiples of the median heuristic. `Median 0.25` refers to a quarter of the median heuristic, while `Median 3` means three times the median, etc. We observed that for the standard t-distribution, the larger the bandwidth, the higher the power of the test. On the contrary, in the case of skewed normal distribution, smaller bandwidths yielded better results in general. This highlights the dependency on distribution of the choice of bandwidth to the power of kernel-based two sample tests.

Given the lack of theoretical foundation to the optimal kernel choice, [34] suggested

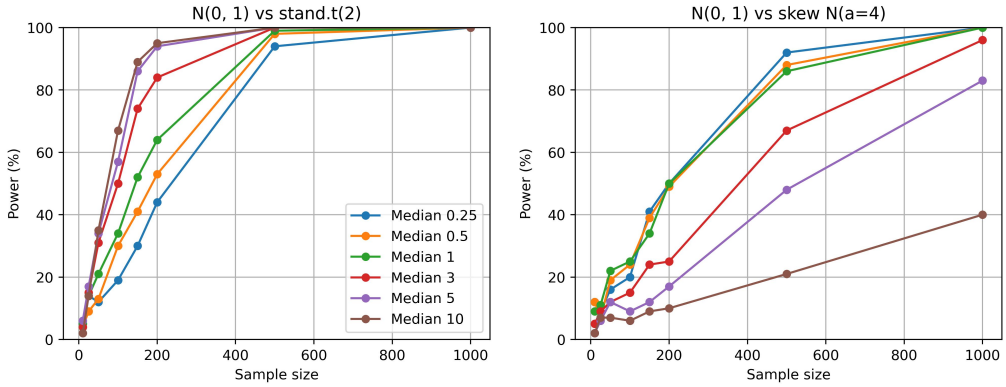


Figure 2.7: Effect of distribution on optimal bandwidth choice at level $\alpha = 5\%$

using multiple kernels to form an aggregated test. This approach enhances the robustness of the test power by utilising a range of bandwidths, and proved to be optimal in the minimax sense. Another strategy, proposed in [18], involves splitting a portion of sample as training data for determining bandwidth that maximises the test power. However, this method reduces the number of samples available for the actual test, which can be problematic in scenarios where minimising sample size is crucial.

As demonstrated, the challenge of choosing an optimal bandwidth is complicated and therefore left as an area for future research. For the purpose of this report, the median heuristic is sufficient to serve as a good baseline for our analysis given its simplicity and widespread use in the literature.

2.1.4 Effect of interpolation

Given the observed relationship between sample size and test power, we investigate whether artificially increasing the sample size through interpolation can improve the power of our two-sample tests. This section explores this proposition, focusing on two interpolation methods: linear interpolation and Brownian bridge interpolation.

Let $(X_i)_{i=1}^n$ be our data sampled, where t_i represents the time. We consider two interpolation methods:

- **Linear Interpolation:** For any $t \in [t_i, t_{i+1}]$, we define

$$X_t = X_i + \frac{t - t_i}{t_{i+1} - t_i}(X_{i+1} - X_i)$$

- **Brownian Bridge Interpolation:** For any $t \in [t_i, t_{i+1}]$, we define

$$X_t = X_i + \frac{t - t_i}{t_{i+1} - t_i}(X_{i+1} - X_i) + \sqrt{\frac{(t_{i+1} - t)(t - t_i)}{t_{i+1} - t_i}}Z$$

  where $Z \sim N(0, \sigma^2)$ and σ^2 is estimated from the original data.

The Brownian bridge interpolation adds a stochastic component to the linear interpolation, potentially better reflects the underlying continuous-time process, especially if we assume the data follows a geometric Brownian motion since the log price follows a drifted Brownian motion.

To evaluate the effect of interpolation on test power, we conducted experiments where we generated geometric Brownian paths $(X_i)_{i=1}^n$ with regime changing at the mid point,

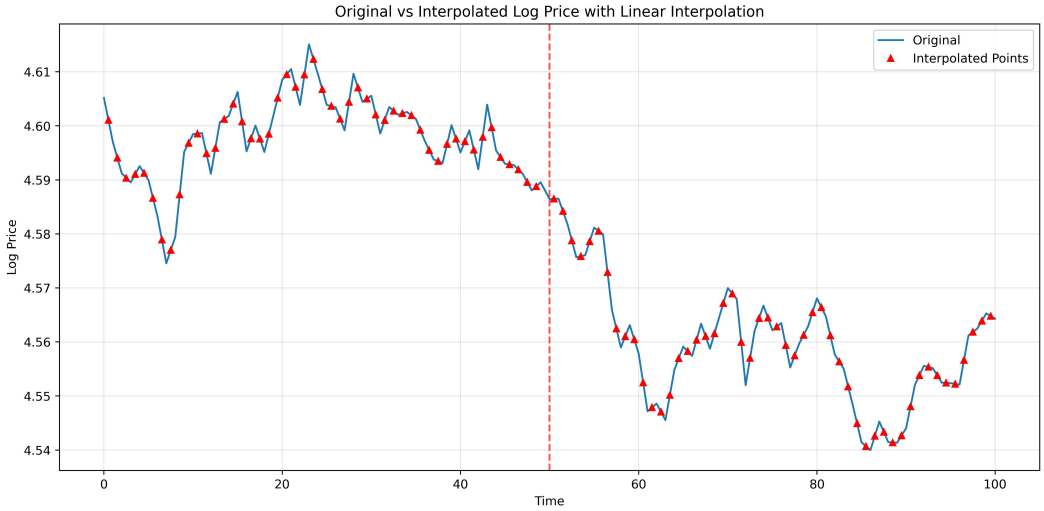


Figure 2.8: Linearly interpolation of log price at mid points

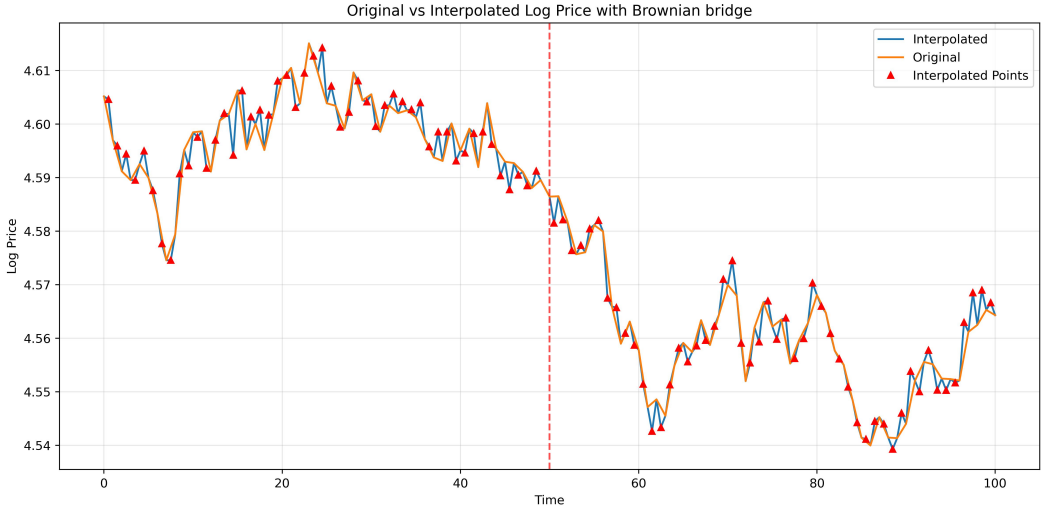


Figure 2.9: Brownian bridge interpolation of log price at mid points

then applied each of the interpolation methods to the log price of $(X_i)_{i=1}^n$. Here we provide the plots of a realised interpolation with the two methods.

As we can see in Figure 2.9, interpolation using Brownian bridge added random fluctuations on top of the linear trend. Therefore, the volatility structure of the Brownian bridge interpolated result resembled actual data more closely. In fact, we empirically estimated the volatility parameter of the interpolated results and saw that interpolation using Brownian bridge recovered almost 70% of the true volatility while linear interpolation only retained 45% of the actual value.

The experimental setup consisted of the following steps: First, we simulated a geometric Brownian motion time series with a change in volatility parameter from 0.2 to 0.25 at the middle of the time series. Each level of interpolation refers to a level of dyadic interpolation; for example, level 1 interpolates the midpoints of the interval, level 2 interpolates the mid points of the level 1 interpolation, resulting in three synthetic samples in each interval, and so on. We then performed MMD two-sample tests on the log returns of the original and interpolated data, separating the data at the middle (the known change point). This process was repeated multiple times to estimate the power of the test under

the alternative hypothesis when the volatility changes and the Type I error rate when volatility is constant.

Our results indicated that linear interpolation has a significant impact on the performance of two-sample tests. Under the alternative hypothesis, we observed an increase in the power for all tests when applied to linearly interpolated data, particularly for smaller original sample sizes. However, this improvement turned out to be an illusion as it was accompanied by a significant inflation of Type I error rates in Figure 2.8.

At a significance level of 0.05, we observed Type I error rates of more than $\sim 30\%$ when linearly interpolation was applied. These rates are substantially higher than the expected 5%, indicating that the tests are becoming unreliable.

The inflation of Type I error rates can be attributed to the artificial reduction in variance introduced by linear interpolation. This reduction in variance narrows the sampling distribution of our test statistics under the null hypothesis. As a result, when we apply our usual critical values, we are more likely to reject the null hypothesis even when it is true. The reduced variance makes small, random differences between samples appear more significant than they actually are, leading to an increase in false positives.

This effect is particularly pronounced in tests sensitive to the overall distribution of the data, such as the Kolmogorov-Smirnov or MMD tests, as the interpolation alters the empirical distribution function in a way that can exaggerate small differences between samples.

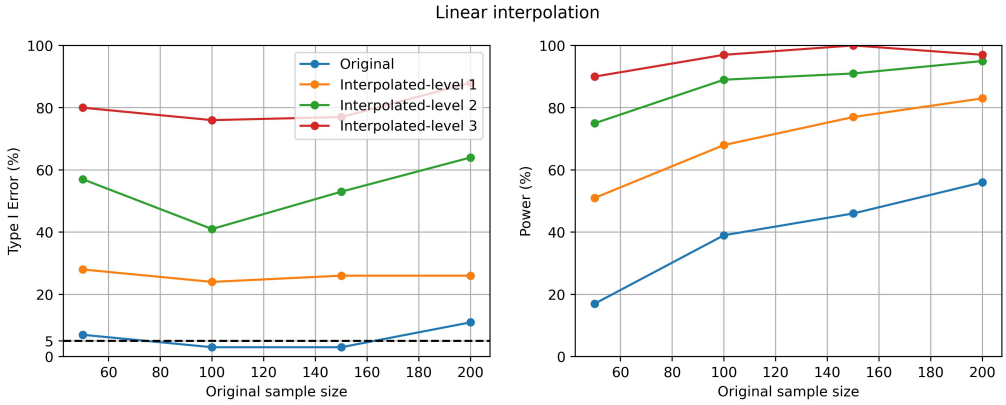


Figure 2.10: Type I error and power (%) for linear interpolation

On the other hand, the results in Figure 2.9 for Brownian bridge interpolation were far superior. Under the alternative hypothesis, we observed a substantial increase in power across all interpolation levels, particularly for smaller original sample sizes. Unlike the case for linear interpolation, the Type I errors were all approximately 5% and aligned with our expectations. The improved performance of Brownian bridge can be attributed to its ability to better preserve the variance structure of the original data by adding a stochastic component.

Notice however, without prior knowledge of the volatility parameter, the value of σ needs to be estimated from the original data, leading to additional bias. Assuming that the underlying process follows a geometric Brownian motion, an estimation of the volatility σ can be obtained by the sample variance of the log return. To illustrate the practical issue of estimation error on the efficacy of the Brownian bridge interpolation scheme we have provided the following plot in Figure 2.12.

The experiment setup was similar to the investigation of Type I error and power under different interpolation schemes. As we are interested in the effect of estimation error, we fixed the sample size of the original generated path and altered the variance parameter of the Brownian bridge. The absolute estimation error is the absolute difference

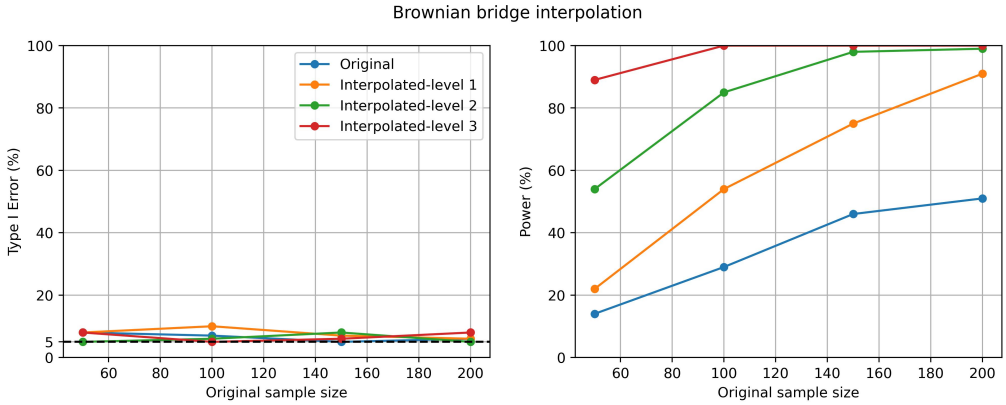


Figure 2.11: Type I error and power (%) for Brownian bridge interpolation

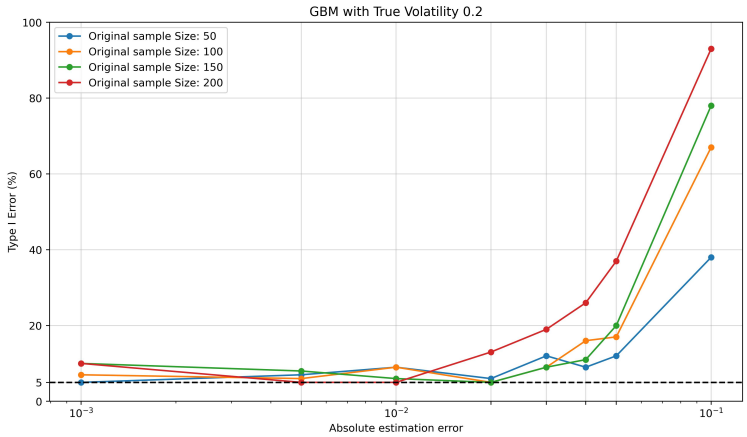


Figure 2.12: Type I error against estimation error of σ under Brownian Bridge interpolation

between the variance parameter used and the true volatility of 0.2. For all fixed sample sizes, the first level of interpolation was applied i.e. mid points of the original data were interpolated. And the x-axis was plotted on a $\log_{10}$ scale to better visualise the relationship across different orders of magnitude. More importantly, we discovered that the Type I error rate increased as the absolute estimation error increased across all fixed original sample sizes. As expected, this phenomenon can be attributed to the fact that inaccuracies in volatility estimation altered the sample distribution to the extent that the two samples being compared were perceived as different. In addition, the rate of increase in Type I error was dependent on the size of the original sample. In particular, smaller original sample sizes exhibited a slower inflation of Type I error rates. This observation suggests that the two-sample test we employed was more forgiving of higher variability or inconsistency in smaller datasets. The lower confidence in rejecting the null hypothesis for smaller samples is reasonable, given the fewer observations available. One might interpret this finding as evidence supporting the efficacy of Brownian bridge as an interpolation scheme, particularly since interpolation is typically more relevant for smaller sample sizes. However, in reality, the standard error of the sample variance estimator is proportional to $\frac{1}{\sqrt{N}}$ with N being the number of observations. This implies that estimation errors are innately larger when working with small sample sets.

Also, the use of Brownian bridge relies on the assumption of normal distributed log returns. In practice, this assumption must be validated before adopting the Brownian bridge approach to ensure theoretical robustness. These findings highlight a crucial point:

while interpolation can increase the apparent sample size, it does not add genuine new information to the dataset. The interpolated points introduce biases that can lead to unreliable conclusions. One should be mindful when considering interpolation as a means to augment sample sizes, and should be aware of the potential inflated Type I error.

2.1.5 Kernel Fisher Discriminant ratio

In the follow section, we investigate another kernel-based two-sample test closely related to the MMD test. In order to define the test statistics, we need to first define a covariance operator.

Definition 2.1.2. (Covariance operator). The covariance operator Σ_p of measure p is the unique linear operator onto $\mathcal{H}$ such that

$$\mathrm{Cov}_p(f(x), g(x)) = \langle f, \Sigma_p g \rangle \qquad \forall f, g \in \mathcal{H}$$

Definition 2.1.3. (Maximum Kernel Fisher discriminant ratio, [12], Section 3.1). Let $\mathcal{H}$ be the RKHS defined in Definition 1.1.2 and two sets of random variables $(x_1, \ldots, x_{n_1}) \overset{i.i.d.}{\sim} p$, $(y_1, \ldots, y_{n_2}) \overset{i.i.d.}{\sim} q$. The maximum kernel Fisher discriminant ratio is defined as

$$\mathrm{KFDR}[\mathcal{H}, p, q] := \frac{n_1 n_2}{n_1 + n_2} \|(\Sigma_W + \gamma I)^{-1/2}(\mu_p - \mu_q)\|_{\mathcal{H}}^2$$

where γ is a positive constant and $\Sigma_W := \frac{n_1}{n_1+n_2}\Sigma_p + \frac{n_2}{n_1+n_2}\Sigma_q$ is the pooled covariance operator.

There are two aspects about kernel Fisher discriminant ratio that are interesting in contrast to MMD. Firstly, the standardised KDFR asymptotically converges to a standard normal distribution under the null hypothesis when the sequence γ_n decreases in a rate slower than $n^{-1/2}$. This result suggests an alternative hypothesis testing procedure. As we have seen in [17], the asymptotic null distribution of MMD converges to an infinite series of chi squared distribution, thus we can only obtain the p-value by bootstrapping which increases computation cost. Therefore, convergence to a simple analytical distribution would be an immense improvement. We highlight that this convergence result is not completely distribution independent. The decaying rate of γ is a function of the sample size n, choice of kernel and distributions p, q. However, we have shown empirically in Figure 2.13 by choosing γ small enough 10^{-5}, the approximation to standard normal is close.

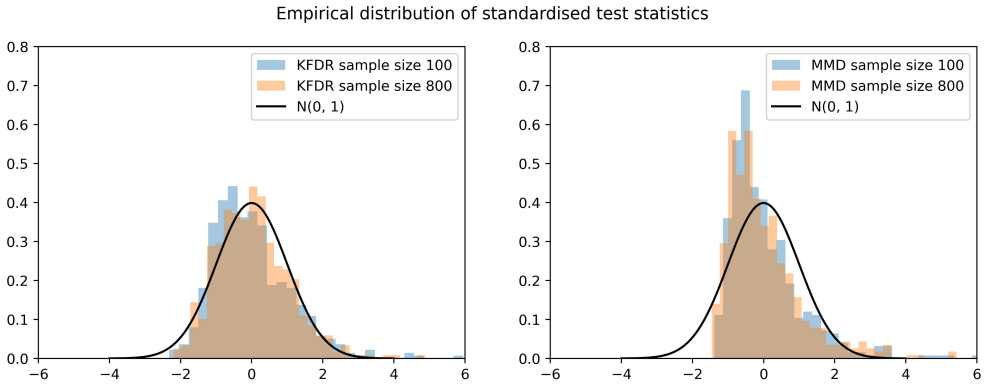


Figure 2.13: Normality of standardised KFDR vs standardised MMD

The second point of interest from [22] is the proposal of a KFDR based two-sample test applicable to unknown change point location. The procedure works as follows. Given

a sequence of samples X among which there exists a point k* where the generating distribution changes, the test statistic is defined as

$$\widehat{T} = \max_k \frac{\widehat{\mathrm{KFDR}}(X[:k], X[k:]) - d_1}{d_2}.$$

where $d_1 = \mathrm{Tr}\left\{\left(\hat{\Sigma}_W + \gamma\,\mathrm{I}\right)^{-1}\hat{\Sigma}_W\right\}$, $d_2 = \mathrm{Tr}\left\{\left(\hat{\Sigma}_W + \gamma\,\mathrm{I}\right)^{-2}\left(\hat{\Sigma}_W\right)^2\right\}$.

The null hypothesis is rejected if T exceeds a certain threshold. This procedure relies on the fact the standardised KFDR peaks at the true change point, hence raising an interesting question as to whether this property holds when there are multiple change points in the time series. A similar idea was explored in [33] which introduces the non-parametric binary segmentation which recursively bisects the time series at indices with maximum Kolmogorov-Smirnov statistics.

To answer the question, we conducted the following test. We randomly split a length 500 time series into four unequal segments. Then, we sampled from a normal distribution with variance alternating between 0.3 and 0.6 for each segment. Here we plot the values of KFDR standardised by its mean and variance empirically estimated with 25 permutations. We also introduce a new hyper-parameter `win size` for the test which governs the scope of the KFDR estimator. This is for two reasons, the first being that the computation of KFDR for $X[:k]$ and $X[k:]$ can be very costly when the time series is long as it has a complexity of $\mathcal{O}(n^3)$. Therefore, limiting the size of the test to `win size` ensures consistency in computational performance. Also, we noticed that the choice of the `win size` parameter impacts the prominence of the peaks. When the window size is too small, the KFDR may be too sensitive to local fluctuations in the data, leading to multiple spurious peaks that do not correspond to true change points. On the other hand, when the window size is too large, the KFDR may be overly smoothed, causing the peaks at the true change points to become less pronounced.

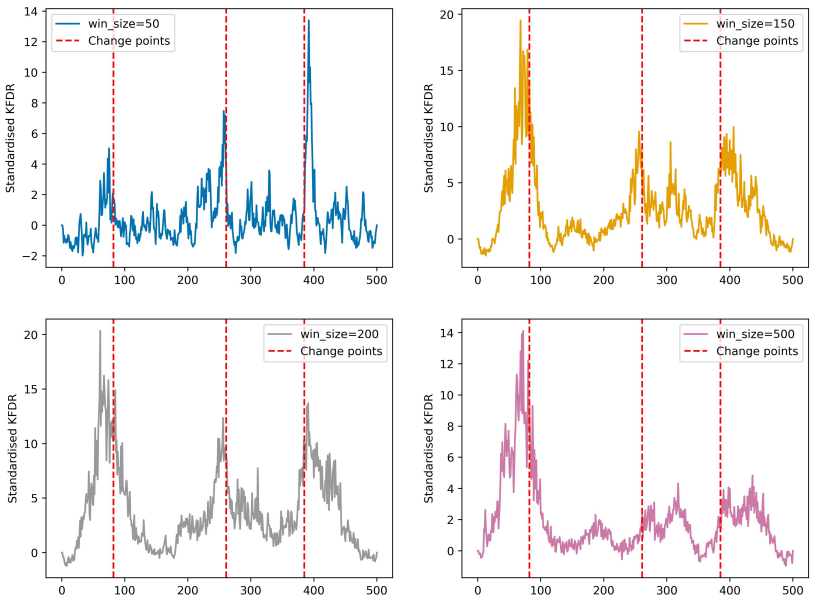


Figure 2.14: Rolling KFDR with varying window sizes

Notably, in Figure 2.14 the KFDR statistic is shown to exhibit peaks at the location

of change points, confirming the test's ability to identify structural shifts in time series, even in the presence of multiple transitions. The plots also demonstrate the effect of the parameter `win size`. As hypothesized, a small window size exhibits noisier values while a large window size is less able to capture changes. Hence, the optimal range of `win size` values should be chosen with care to strike a balance between sensitivity and robustness when applied to change point detection problems.

The intriguing results observed with the standardised KFDR prompt the question whether other non-parametric tests exhibit similar behaviours. We carried out the same procedure using MMD statistics instead of KFDR on the same synthetic time series. And the plot of varying window size is provided in Figure 2.16 below.

Similar to the KFDR-based analysis, change points in the time series correspond to a peak in the MMD, suggesting that both non-parametric tests can potentially be used to identify candidates of change points. The overall effect of window size is similar to the observation in KFDR as well, with spurious and lower peaks for smaller window sizes and dampened peaks at some change points for larger window sizes.

Lastly, we compared the power of the MMD test with the KFDR test under various distribution shift scenarios, with a particular focus on the skewed generalized t-distributions. Figure 2.15 illustrates the power of both tests under different sample sizes of the skewed generalized t-distribution.

As evident in Figure 2.15, both tests approached a power of 1 as the number of samples increased. While the MMD test performed slightly better for this particular choice of parameter values, KFDR showed a slight edge in performance in the standard normal against student t-test. It is crucial to emphasize that the power of the test is subject to kernel and hyper-parameter choices. Hence, we concluded that the two tests demonstrated comparable power. We provide additional results of the comparison in Appendix B, along with a plot of Type I error to showcase the validity of both tests.

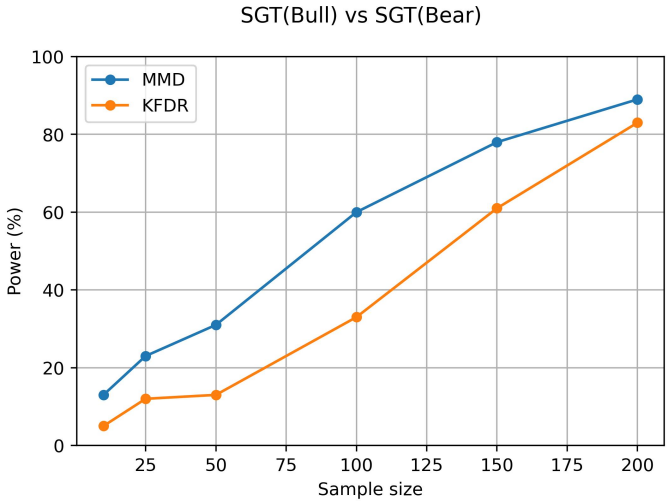


Figure 2.15: Power (%) against sample size at level $\alpha = 5\%$

2.2 Algorithm

In the previous sections, we investigated the performance of two-sample tests and found that kernel-based methods and the Epps-Singleton test stood out among the others for most cases. In the problem of change point detection, the number and location of change points are unknown. Therefore, a two-sample test alone is not enough. Here we propose three change point detection algorithms based on two sample tests.

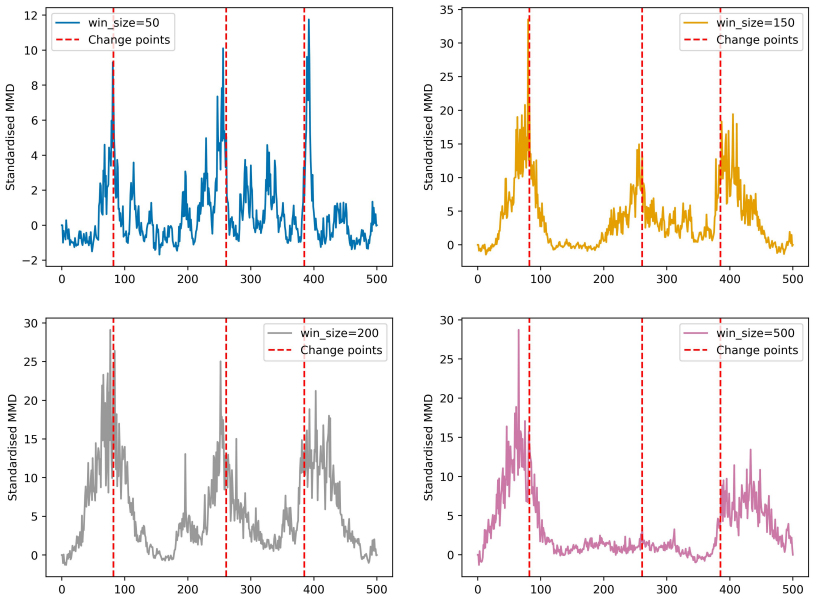


Figure 2.16: Rolling MMD with varying window sizes

The first algorithm we propose is based on a sequence of two sample tests in a sliding window basis. By iterating through each position i in the time series and performing a two-sample test `Test` on window centred at i, we add the position i to the list of candidate change points if the p-value is less than the significance level. Since the distributional difference at points near the true change point is large, points to the left and right of the true change points are often classified as candidates. To obtain the estimated location of the true change points, we first split the candidate points into contiguous subsegments. And return the medians of the subsegments as detected change points. It is worth noting that the choice of window size `win size` is an important consideration. A smaller window size allows for detection of more localized changes but is more susceptible to Type II error, while a larger window size provides more accurate results but may miss closely

Algorithm 1: Sliding Window

Input : Time series, window size (winsize), significance level (siglevel)
Output: Change points
candidate $\leftarrow$ [·];
for $i = 2$ **to** *length(data)* **do**
  left $\leftarrow$ data[i − winsize : i];
  right $\leftarrow$ data[i : i + winsize];
  pvalue $\leftarrow$ `Test(`*left, right*`)`;
  if *pvalue* $\leq$ *siglevel* **then**
    Append i to candidate;
  end
end
subsegments $\leftarrow$ `Split(`*candidate*`)` to continguous sbusegments;
return `Median(`*subsegments*`)`;

spaced change points. The complexity of the algorithm is $\mathcal{O}(N)$ where N is the length of the time series. Therefore, in the case of the MMD two sample test, the total complexity would be $\mathcal{O}(N \cdot W^2 \cdot P)$ where W is the window size and P is the number of permutations.

Algorithm 2: Binary Segmentation

```
Input  : Time series, minimum segment size (min size), significance level (sig
         level)
Output: Change points
Function BinSeg(data):
    m ← mid point;
    N ← Length(data);
    leftseg ← data[:m];
    rightseg ← data[m:];
    if N ≤ min size then
        p value ← Test(leftseg, rightseg);
        if p value ≤ sig level then
            return [m];
        else
            return [];
        end
    else
        left change points ← BinSeg(leftseg);
        right change points ← BinSeg(rightseg);
        left end ← last element of left change points;
        right end ← first element of right change points;
        p value ← Test(data[left end: m], data[m:right end]);
        if p value ≤ sig level then
            return [m] + left change points + right change points;
        else
            return left change points + right change points;
        end
    end
```

The binary segmentation algorithm recursively bisects the time series into increasingly smaller segments and tests for change points in a bottom up fashion. When the segment length is below a specified minimum size, the algorithm performs a two-sample test and returns the mid point as a change point if the p-value is below the significance level. For progressively larger segments, an additional test is performed on subsegments to the left and right up to the closest change point being identified. The mid point of this larger segment is again included if the test is significant. The recursive depth of the algorithm is controlled by the parameter `min size`. While a smaller `min size` increases the chance of a change point lying on the edge of the segment, the lower power hinders the ability to detect differences in distribution. Note that this algorithm has complexity $\mathcal{O}(\frac{N}{m})$ where m is the minimum segment size.

Lastly, we generalised the analysis in standardised kernel-based statistics to obtain an algorithm for multiple change point detection. Similar to the sliding window approach, the standardised kernel-based statistic is calculated at each position `i` of the time series. In order to identify significant peaks that serve as candidates of change points, a smoothing spline is applied to reduce noise in the signal. Then using a peak-finding algorithm to identify local maxima in the smoothed data, local maxima are returned as potential change points. This procedure introduces a smoothing parameter that governs the smoothness of

Algorithm 3: Standardised Maximum Partitioning

Input : Time series, window size (winsize)
Output: Change points
N ← `Length(`*data*`)`;
standardised stat ← [·];
for $i = 2$ **to** `Length(`*data*`)` **do**
  left ← data$[i - \text{winsize} : i]$;
  right ← data$[i : i + \text{winsize}]$;
  stat ← `TestStatistic(`*left, right*`)`;
  mean, std dev ← `Bootstrap(`*left, right*`)`;
  Append $\frac{\text{stat}-\text{mean}}{\text{std dev}}$ to standardised stat;
end
smoothed data ← `SmoothingSpline(`*standardised stat*`)`;
peaks ← `PeakFinder(`*smoothed data*`)`;
return peaks;

the standardised statistics. However, we want to stress the purpose of smoothing is not to remove insignificant peaks but rather a solution to finding peaks for discontinuous data where all points are essentially local maxima. To ensure robustness in statistical sense, we apply the two-sample test procedure on segments defined by the potential change points. This verification process has the advantage of better defined segments compared to the Binary Segmentation and Sliding Window algorithm where segments of arbitrary length are tested with mid points assumed to be the locations of change points.

As we have mentioned in previous sections, the computation cost grows with the length of time series and the window size. To parallelise the MMD and KFDR calculation, we generalised the implementation into taking a matrix of arbitrary dimension as input. This allows the calculation of p permutation of window size n samples for b points in the time series where b is the batch size to be formulated as a 4D matrix with dimension $(p \times b \times n \times n)$. We then use the Cupy library to parallelise the operation on a GPU. We observed a 4-fold improvement in computation speed. It is interesting to note that the batch size should be carefully chosen so that the GPU's memory is not exceeded as then memory would be borrowed from the CPU, resulting in slower transfer of data. We provide a plot of computation speed up against batch sizes to illustrate the idea in Appendix B.

2.3 Synthetic Data

To systematically evaluate the performance of the algorithm on real data is near impossible. It is very difficult to infer from a sequence of log returns the underlying probability distribution each segment originates from. Nor is it possible to tell the exact locations of change points should a change in distribution occurs. As a result, we tested our algorithms on synthetic market data generated from either a geometric Brownian motion or Merton jump diffusion process.

For a given interval $[0, T]$ with $T \in \mathbb{N}$, we generated time series with timestep Δt, resulting in a total number of $N = \frac{T}{\Delta t}$ sample points. Let $\Theta \subset \mathbb{R}^d$ be the parameter space and $(\theta_1, \ldots, \theta_m) \in \Theta^m$ be a set of parameters, each specifying the path generation model. For the purpose of this analysis, $m = 2$, with parameter set labelled $(\theta_{\text{bull}}, \theta_{\text{bear}})$ corresponding to the parameters of bull and bear market regime. We then drew K random samples from the index set $[0, \ldots, N]$ as the locations of change points, with K being the pre-specified number of change points. To prevent consecutive change points, we introduced the parameter `min size` to enforce a minimum interval size between any two

consecutive change points. Starting from the first element in the set of parameters, we simulated a path associated with the parameter. When a change point occurs, we cycled to the next item in the set of parameters, and simulated the corresponding path until the next change point was encountered.

To assess the performance of various change point detection algorithms, we employed three key metrics: Hausdorff distance, Rand index, and F_1 score.

- **Hausdorff Distance** The Hausdorff distance measures the maximum distance between the detected change points and the true change points. It quantifies how far the estimated change points are from the actual ones in the worst case. Thus it provides a measure of the worst-case error in change point location, assessing the maximum deviation of the algorithm. A smaller Hausdorff distance indicates better performance.
$$d_H = \max\{\sup_{a\in A}\inf_{b\in B} d(a,b), \sup_{b\in B}\inf_{a\in A} d(a,b)\}$$

- **Rand index** The Rand index evaluates the similarity between the segmentation produced by the estimated change points and the true segmentation. It considers all pairs of time points and checks if they are correctly classified as being in the same segment or different segments. As a result, it measures the overall structural similarity by considering the entire segmentation instead of individual change points. The Rand index ranges from 0 to 1, with 1 indicating perfect agreement. This is beneficial for assessing overall structural similarity.
$$\text{RI} = \frac{TP + TN}{TP + FP + TN + FN}$$
where TP, TN, FP, and FN are true positives, true negatives, false positives, and false negatives, respectively.

- **F_1 Score** The F_1 score is the harmonic mean of precision and recall.
$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$
It ranges from 0 to 1, with 1 indicating perfect precision and recall. The F_1 score penalises false positives as well as false negatives. Thus it is a robust measure of the algorithm's accuracy.

2.3.1 Geometric Brownian motion

In the following subsection, we discuss the test results of the algorithms on identifying change points under the geometric Brownian motion model. The parameter set θ that determines the process is given by (μ, σ) where μ is the drift coefficient and σ the volatility coefficient. As a result, the log return of the synthetic price data is given by

$$\Delta \log S_t \sim N((\mu - \frac{\sigma^2}{2})\Delta t, \sigma^2 \Delta t).$$

We simulated a path generated by the set of parameter values provided below,

$$\theta_{\text{bull}} = (0.04, 0.2),$$

and

$$\theta_{\text{bear}} = (-0.04, 0.3).$$

We provide below Table 2.1 of the mean and standard deviation performance metric estimates over 50 simulated paths.

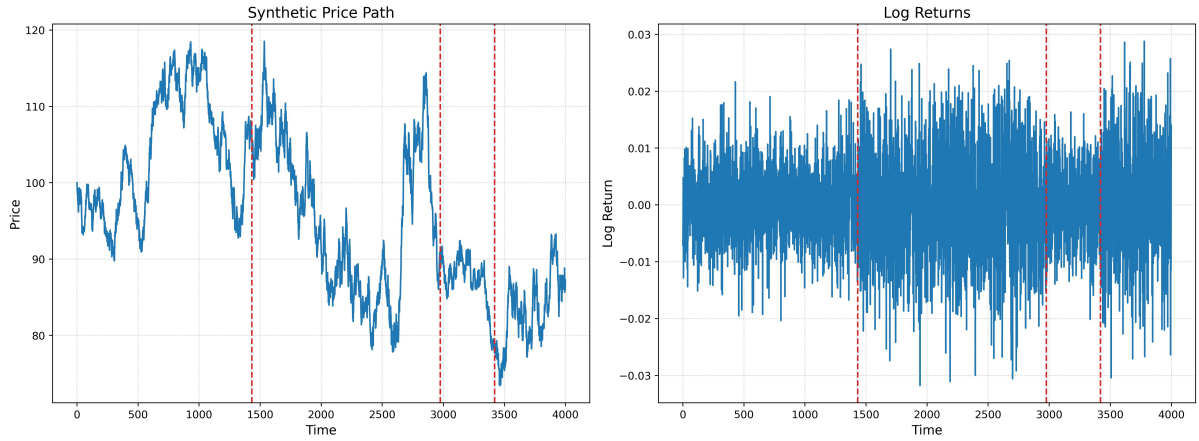


Figure 2.17: Synthetic path under geometric Brownian motion

Method	Hausdorff	Randindex	F_1 score
SMP MMD	332.96 ± 456.69	0.94 ± 0.07	0.84 ± 0.18
SMP KFDR	258.56 ± 450.61	0.96 ± 0.08	0.94 ± 0.10
Binseg MMD	538.96 ± 485.70	0.91 ± 0.06	0.74 ± 0.16
Binseg ES	609.28 ± 472.16	0.89 ± 0.09	0.69 ± 0.14
Win MMD	1048.92 ± 556.98	0.90 ± 0.07	0.71 ± 0.11
Win ES	1302.62 ± 459.28	0.88 ± 0.09	0.61 ± 0.11

Table 2.1: Performance metrics for change point detection algorithms under Geometric Brownian Motion

The results indicate that the SMP KFDR method outperformed other algorithms across all metrics. It achieved the lowest average Hausdorff distance of 258. This means that the furthest distance between an estimated change point and its corresponding true location is very small, showing that the method is robust. The close to one Randindex indicated almost perfect overlap between the true and estimated segmentation. Finally, the highest F_1 score suggests that SMP KFDR not only correctly identified change point locations, but rarely misclassified a non change point.

The SMP MMD method showed the second-best performance across all metrics, with a notably low Hausdorff distance of 332, high Randindex of 0.94 and F_1 score of 0.84. This shows that SMP MMD is also highly effective in detecting change points under geometric Brownian motion.

Binary segmentation methods showed moderate performance, with Binseg MMD slightly outperforming Binseg ES across all metrics. However, their Hausdorff distances were significantly higher than those of the SMP methods, suggesting less precise change point detection. The lower F_1 score might be due to the higher number of estimated change points than there actually were. Even though true change points were correctly identified as suggested by the close to 0.9 Randindex, the increased false positive dragged down the F_1 score. The lower Randindex compared to SMP is expected though. The binary segmentation inherits two limitations as we have briefly discussed in the previous section. First, to achieve higher accuracy of the estimated change point location, we chose a relatively small `min size` of 100. This compromised the power of the test. Also, the performance of the binary segmentation is subject to the location of the true change points. When a change point lies in the middle of the `min size` wide window, the test rarely has enough evidence to reject the null hypothesis. These shortcomings suggest that binary segmentation might not be the right algorithm to use when investigating real-world data.

The window-based methods demonstrated the worst performance, particularly in terms of Hausdorff distance. By comparing the set of estimated change points and true change points, we observed that window based methods were much noisier than the previous two algorithms, meaning that the number of false positive is high. In a lot of instances, change points were misidentified near the start and the end of the time series. This might be due to the imbalanced number of samples when two sample tests were carried out at the beginning and the end of the time series. As evidenced in [4] and previous sections, we showed that the small number of total samples together with large differences between the two sample sets reduce the power of the two-sample test. With only a few points at the two ends of the time series to test on, misclassified points inflated drastically the Hausdorff distance. Other than this, the performance was similar to the binary segmentation. And again, Win MMD slightly outperformed Win ES.

2.3.2 Merton jump diffusion process

In the previous case, the log return of the price series is Gaussian distributed. In order to capture discontinuity behaviours in the market, we investigate the performance under a Merton jump diffusion process. The solution to the stochastic differential equation is given by

$$S_t = S_0 \exp\left((\mu - \frac{\sigma^2}{2})t + \sigma W_t + \sum_{i=1}^{N_t} Y_i\right)$$

where W_t is Brownian motion, N_t a Poisson process with intensity λ and $(Y_i)_{i=0}^n$ a sequence of i.i.d. normal random variables with mean m and variance v.

The sets of parameters corresponding to the bull and bear market we used are given by following,

$$\theta_{\text{bull}} = (0.05, 0.2, 5, 0.02, 0.0125),$$

and

$$\theta_{\text{bear}} = (-0.05, 0.4, 10, -0.04, 0.1).$$

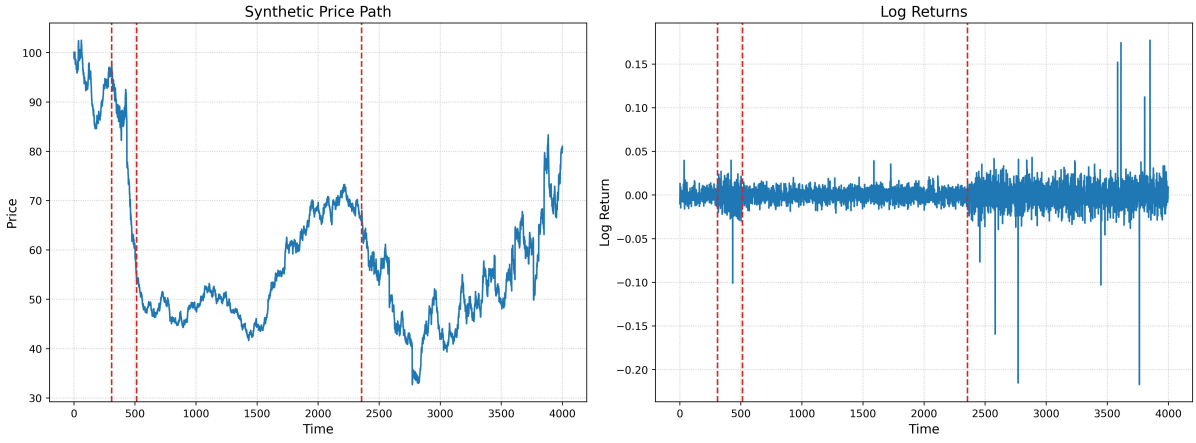


Figure 2.18: Synthetic path under Merton jump diffusion

Again, 50 synthetic paths were drawn to estimate the mean and standard deviation of the performance metrics. And the results are presented in Table 2.2

In the Merton jump diffusion scenario, we observed that SMP KFDR maintained its superior performance, achieving the best results across all metrics. It showed the lowest Hausdorff distance of 371, highest Randindex of 0.96, and best F_1 score of 0.89.

Method	Hausdorff	Randindex	F_1 score
SMP MMD	682.72 ± 734.74	0.91 ± 0.10	0.77 ± 0.18
SMP KFDR	371.34 ± 648.47	0.96 ± 0.08	0.89 ± 0.14
Binseg MMD	636.30 ± 536.32	0.90 ± 0.08	0.74 ± 0.15
Binseg ES	654.10 ± 533.44	0.89 ± 0.08	0.67 ± 0.13
Win MMD	909.98 ± 537.80	0.88 ± 0.09	0.73 ± 0.13
Win ES	1387.44 ± 615.55	0.88 ± 0.12	0.64 ± 0.11

Table 2.2: Performance metrics for change point detection algorithms under Merton Jump Diffusion Process

SMP MMD, while still performing well, showed a noticeable decrease in performance compared to the geometric Brownian motion case, particularly in terms of Hausdorff distance. This suggests that SMP MMD is more sensitive to the discontinuities introduced by the jump process. Therefore, it wrongly identified jumps as changes in distribution.

The binary segmentation methods showed very similar performance to the geometric Brownian motion case, with Binseg MMD slightly outperforming Binseg ES. Their performance was closer to that of the SMP methods in this scenario, indicating that they may be more robust to jumps than initially thought.

Window-based methods again showed the poorest performance, with larger Hausdorff distances compared to the geometric Brownian motion scenario. However, the Randindex and F_1 score were similar to that of the geometric Brownian motion, demonstrating the resilience in ability to accurately identify change points in discontinuous time series.

Overall, the introduction of jumps in the Merton model increased the difficulty of change point detection for all methods, as evidenced by the generally higher Hausdorff distances. However, SMP KFDR demonstrated remarkable robustness, maintaining high performance across both scenarios. The consistent superior performance of SMP KFDR across both scenarios indicates its versatility in handling different types of price dynamics where there might be both continuous changes and sudden jumps. Other methods also exhibited moderate robustness in selecting the correct change points.

Chapter 3

Application: Prediction

Building upon the foundation of offline change point detection methods discussed in the previous chapter, we turn our attention to the more challenging and practically relevant task of early detection and prediction of regime changes in financial markets. While data before and after change points are available to offline approaches, real-world applications often require timely identification of regime shifts.

In this chapter, we first explore the idea of early detection in synthetic data that predicts regime changes before they fully manifest with tools we developed. By studying the efficacy and sensitivity of the tests under various known dynamics, we extend the application using machine learning techniques to real market data, in the hope of creating models capable of anticipating significant structural breaks and offering practical tools for strategy risk management.

3.1 Synthetic example

In synthetic data, regimes are generated independent to one another, which inherently precludes any predictability before a change point occurs. To introduce an element of predictability into our synthetic examples, we incorporate a transition period during which the regime gradually evolves towards the next state. This modification allows us to explore the potential for early detection of regime shifts.

In our study, we labelled the 50 time points immediately preceding a complete regime shift as '1', indicating the pre-change period of interest. The remaining points were labeled as '0'. For our predictive model, we utilized the rolling MMD statistics as the primary independent variable. We also included three lagged versions of this statistic to capture the trend of the statistics, each separated by a step size of 100 points. This results in a total of four independent variables for our classification task.

We employed logistic regression as our classification model to determine whether a given data point is on the edge of a regime shift based on the evolving statistical properties of the time series during the transition period. Our synthetic dataset comprised of 200,000 time steps with 100 change points, each separated by a minimum of 1,000 time steps. This configuration resulted in a significant class imbalance, with a ratio of approximately 1:40 between the '1' labels and '0' labels. To address this imbalance, we implemented a weighting scheme where data points were assigned weights inversely proportional to their class frequency, ensuring that the model adequately learns to identify the crucial pre-change periods despite their relative scarcity in the dataset.

The performance of our logistic regression model was evaluated under three distinct market dynamics: Geometric Brownian motion, Merton-Jump Diffusion and Skewed Generalised t-distribution. These models represent varying levels of complexity in the bull and bear market return. To ensure a comparative assessment of model performance, we

employed SMOTE (Synthetic Minority Over-sampling Technique) [8].Instead of directly computing the F_1 score on the testing data, SMOTE addresses the class imbalance issue by randomly interpolating between similar minority data points to create new synthetic samples. This prevents potential biases in the F_1 score calculation, such as inflated scores from simply classifying all points as label 0 or deflated scores due to a relatively high number of false positives.

For the GBM model, we observed a generally increasing trend in F_1 score as the transition period lengthens. The F_1 score ranges from 0.66 to 0.78, indicating moderate predictive performance that improved with longer transition periods.

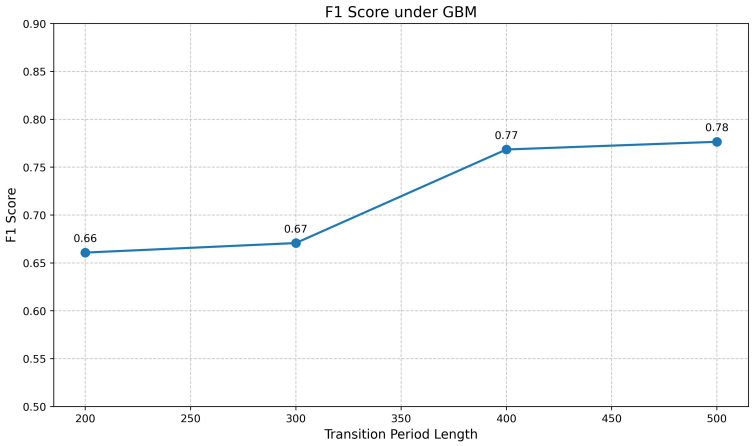


Figure 3.1: F_1 score against length of transition period under GBM

The MJD model, which incorporates sudden jumps, showed the highest overall F_1 scores among the three dynamics. The scores range from 0.7 to 0.89, with a sharp increase up to 400 time steps and then a slight dip at 500 time steps.

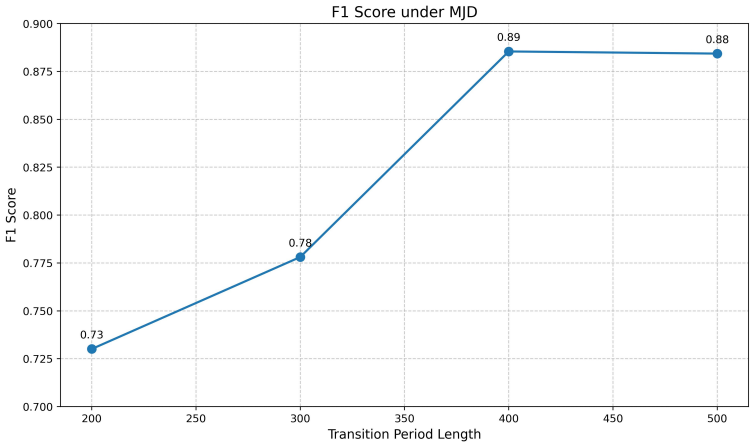


Figure 3.2: F_1 score against length of transition period under MJD

Lastly, using the same parameters as in the power test for Skewed Generalised t-distribution, the SGT model exhibited a steady increase in F_1 score from 0.63 to 0.73 as the transition period lengthened. While these scores were lower than those of the other two models, the consistent improvement with longer transition periods confirms the validity of our approach and demonstrates the ability to detect regime changes, albeit with less identifiable distribution difference as shown in the power test results.

These results demonstrate that the effectiveness of our logistic regression model in predicting regime changes is influenced by both the magnitude of the underlying market

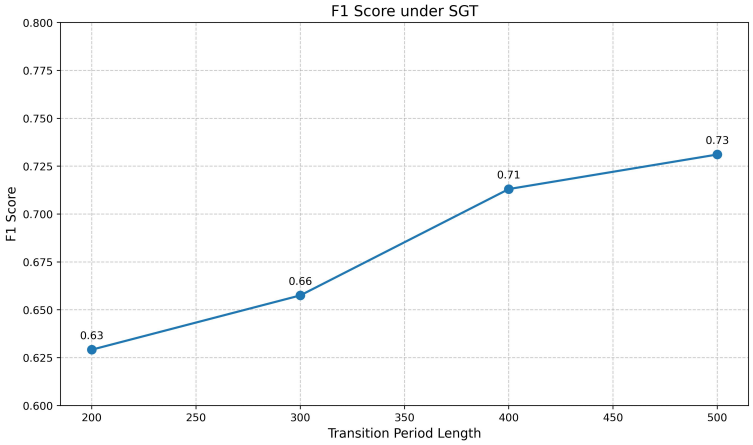


Figure 3.3: F_1 score against length of transition period under SGT

dynamic changes and the length of the transition period. In general the model performs best with Merton-Jump Diffusion dynamics and longer transition periods, suggesting its potential utility in markets characterized by occasional sharp movements and gradual regime transitions. To further assess the robustness of the test results, we carried out the same experiments with a 50/50 train-test split, instead of the 70/30 split presented in the figures above. The resulting F_1 scores were found to be almost identical to the original findings, indicating that there was no issue of overfitting in the models.

In practice, predictions before the change point are more valuable than those after, as they allow for timely adjustments to trading strategies. To investigate this issue, we explored the introduction of asymmetry in the model's cost function. One proposed method involves lowering the weighting of points immediately preceding the label '1' and increasing the weight of points shortly after. This approach emphasizes correctly classifying points after a change point as not being a change point, while reducing the penalty for misclassifying label '0' points near but before change points. Our study revealed that maintaining the same level of true positive accuracy requires an asymmetric trade-off: doubling the weight after a change point required more than halving the weights of points before the change point. This weighting scheme can be tailored to user preferences and strategy focus. We provide the results of an example of this implementation in Appendix B. A potential refinement to explore is the implementation of gradually diminishing weights after the change point and gradually increasing weights for points further away from the change point in the preceding period. This nuanced approach could potentially enhance the model's ability to provide early warnings without sacrificing overall accuracy.

3.1.1 Signature features

While the rolling MMD statistics have shown decent promise in detecting regime changes early, we explore an alternative approach which is not distribution based. The *path signature* provides a rich set of features that capture deep geometric properties of a path. In the following subsection, we briefly discuss important properties of path signatures that are of interest and provide the experiment results of our study.

Definition 3.1.1. (Signature, [9], Definition 1). The signature of a path $X \in C_p([a,b],V)$, the space of continuous path from the interval $[a,b]$ to a d-dimension Banach space V with finite p-variation, denoted by $S(X)$ is the collection of all the iterated integrals of X, i.e.

$$S(X) = (1, S(X)^1, S(X)^d, S(X)^{1,1}, S(X)^{1,2}, \dots)$$

where the superscripts run along the set of all multi-indexes

$$W = (i_1, \dots, i_k | k \geq 1, i_1, \dots, i_k \in \{1, \dots, d\})$$

and the iterated integrals are defined as

$$S(X)^{i_1,\dots,i_k} = \int_{a<t_k<b} \cdots \int_{a<t_1<t_2} dX_{t_1}^{i_1} \dots dX_{t_k}^{i_k}.$$

There are three main properties of signatures that make them particularly useful in machine learning. The first is that the magnitude of the signature terms decays at a rate proportional to the factorial of their level.

Proposition 3.1.2. *(Factorial decay, [6], Proposition 1.2.3). Let X be a path in $C_1([a,b], V)$. Then for any $k \in \mathbb{N}$,*

$$\|S(X)^{(k)}\|_{V^{\otimes k}} \leq \frac{\|X\|_{1,[a,b]}^k}{k!}$$

where $\|X\|_{1,[a,b]}^k$ is the 1-variation of X.

As a result of this proposition, higher order terms become increasingly less significant. Therefore, we can use a truncated signature to effectively represent the infinite-length collection of signature terms. Another key property is that the signature characterizes the corresponding path uniquely, up to tree-like equivalence.

Theorem 3.1.3. *(Uniqueness, [6], Theorem 1.4.1). Let $1 \leq p < 2$, then tree-like equivalence $\sim$ defines an equivalence relation on $C_{0,p}(V)$, the subspace of $C_p(V)$ with paths starting at the zero vector $0 \in V$. Moreover, it coincides with the equivalence relation defined by the equality of signatures, i.e. let $X, Y \in C_{0,p}(V)$*

$$X \sim Y \iff S(X) = S(Y).$$

Lastly, to justify the use of signature features in a linear machine learning model, we have the following result.

Theorem 3.1.4. *(Universal approximation with signatures, [6], Theorem 1.4.7). Let $1 \leq p < 2$ and assume K is a compact subspace of C_p. Let $C(K)$ be the space of continuous functions on K with the topology of uniform convergence. If the set $\mathcal{A}_K := \{\Phi_{f|K} : f \in T((V))^*\}$ is a subset of $C(K)$, then it is a dense subset. $\Phi_{f|K}$ denotes the restriction of function Φ_f to K, where*

$$\Phi_f : [X] \to (f, S(X)).$$

To evaluate the effectiveness of signature features for change point detection, we employed the following experimental setup: we generated synthetic time series data with known regime changes, similar to our previous MMD-based experiments. Then for each time point t, we calculated the signature of the path using a sliding window approach with a window size of 200 and a maximum signature truncation level of 6. To better capture the quadratic variation of the path, we augmented the one-dimensional path with a lead-lag transformation, see [9] Section 2.1.2. Additionally, as the magnitude of signature terms decay at a rate of $k!$ (where k is the level), we applied rescaling to facilitate model fitting. Both of these operations were carried out using `esig` Python library. To improve results, we also included lagged versions of the signature terms, similarly to the MMD experiment setup. Finally, all features were standardised before fitting a logistic regression, ensuring consistent scale across variables.

Under the GBM model, the highest overall F_1 scores were achieved. Starting at 0.89 for level 2, it reached 0.94 at level 6. The improvement was more gradual compared to

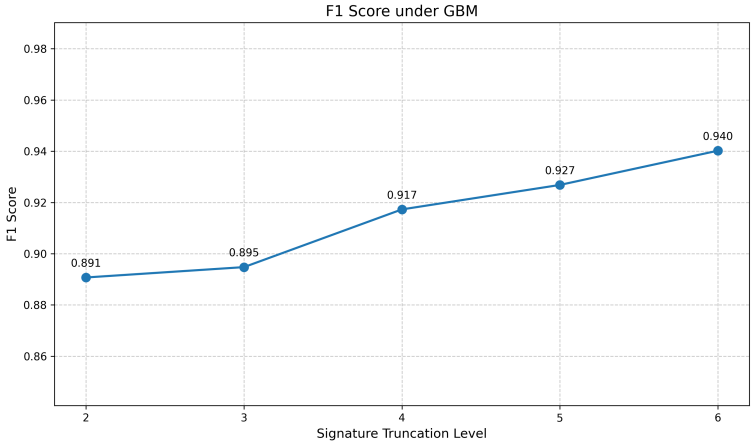


Figure 3.4: F_1 score against truncation level under GBM

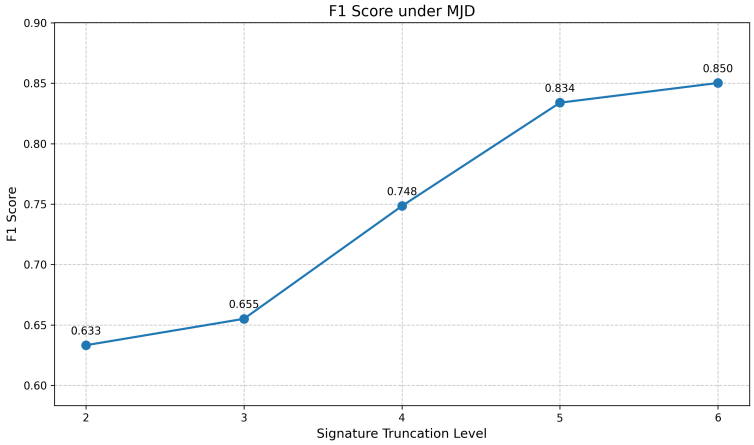


Figure 3.5: F_1 score against truncation level under MJD

other models, suggesting that lower-order signature terms contain enough information to identify a change point in GBM process. This aligns with our expectation since the two alternating regimes in the GBM model differs mainly in volatility.

Interestingly, for the MJD model, we observed a consistent increase in F_1 score as the truncation level increases. The F_1 score rose from 0.63 at level 2 to 0.85 at level 6, indicating that higher order signature terms contributed significantly to characterising abrupt changes and jumps in the time series.

Lastly, for the SGT model, we saw a more pronounced improvement in F_1 score as the truncation level increased. The score started at 0.79 for level 2 and reached 0.92 at level 6. The most substantial jump occurs between level 3 and 4, implying that the fourth order signature terms were particularly useful in detecting change points in the SGT process.

Across all models, we observed that increasing the signature truncation level consistently improved the F_1 score. This improvement was particularly evident when lead-lag transformation was applied to the data, demonstrating that signature features are effective in characterising stochastic paths for predictive learning. However, we are interested in whether this predictive power can be generalised to a broader mixture of processes. While overfitting is unlikely given our experiments with various train-test splits, it is possible that this signature approach captures the characteristics of the path during the transition period rather than identifying a switch in regime. Hence, despite these promising results, further investigation is required to apply signature-based methods in real-life regime switch

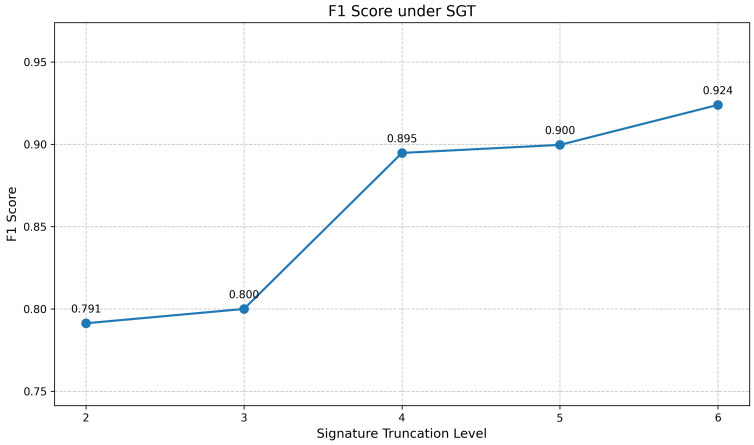


Figure 3.6: F_1 score against truncation level under SGT

detection.

3.2 Real Data

Building upon our insights from the synthetic experiment, we extended our investigation to real-world market data, with an aim of assessing the efficacy of our change point detection model in a more realistic setting. For this study, we utilised the daily price of S&P500 index spanning from January 1994 to August 2024. Similar to our synthetic example, we employed a rolling window approach to compute the MMD statistics and signature, together with their lagged versions as predictor variables.

We maintained the same labelling strategy, marking the 50 trading days preceding identified change points as '1' and the rest as '0'. Change points in the real data were determined using the SMP algorithm with MMD statistics outlined in the offline detection section. The identified change points are presented in Figure 3.7.

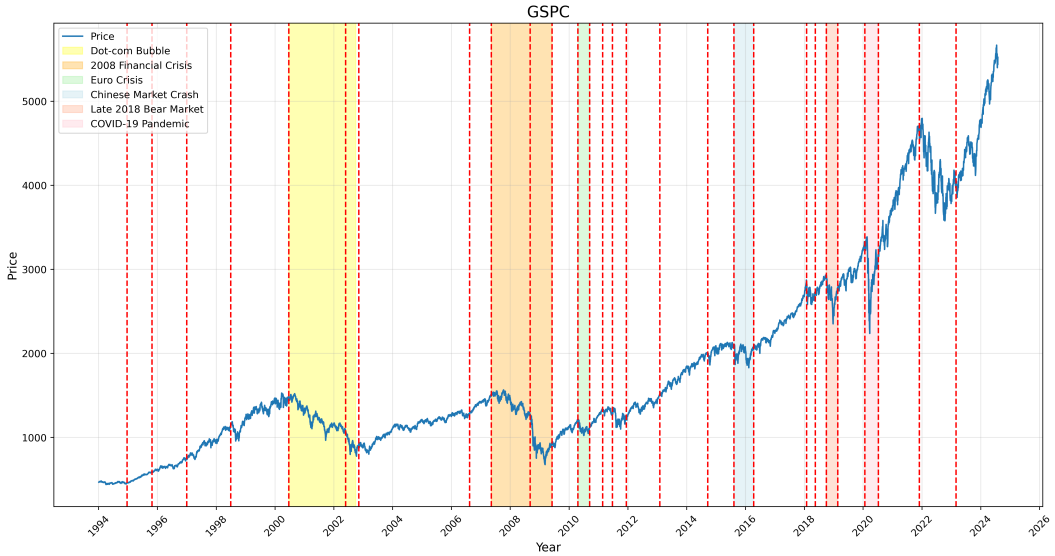


Figure 3.7: Estimated change points

Upon applying our logistic regression model to real-world data, we noticed massive reduction in F_1 score across both feature sets compared to synthetic data.

Feature	In-Sample F_1 Score	Out-of-Sample F_1 Score
MMD	0.37	0.32
Signature	0.77	0.49

Table 3.1: In-sample and out-of-sample F_1 score

The poor F1 scores observed in Table 3.1 for our real-world data application highlight significant challenges in change point prediction. Firstly, both in-sample and out-of-sample results using MMD are very poor, with F_1 scores of 0.37 and 0.32 respectively. This under-performance revealed two critical issues in our approach to change point prediction. Firstly, our previous assumption about the regime transition period does not appear to hold. Even if a transition period exists, its length would not be in the magnitude of hundreds of days, but likely in days which is hard to capture. Consequently, simply monitoring the rolling MMD statistics of log returns immediately before a change occurs offers little predictive power. Additionally, our method of labelling utilizes the rolling MMD statistics, which are also used as predictive features. This approach inherently includes any transition period within the succeeding regime, restraining our ability to detect pre-change signals.

As for the signature feature set, while the in-sample F1 score is 0.77, the out-of-sample result of 0.49 indicates the model struggled to generalize beyond the training data. This discrepancy can be explained by the difference between our synthetic experiments and real market dynamics. In synthetic data, regime transitions were repetitive, allowing the model to learn patterns that could easily generalize to unseen data. However, in real market scenarios, distribution shifts are not repetitive. For example, the burst of the dot-com bubble differs from the housing bubble burst in 2007. Consequently, the model's ability to fit historical events does not necessarily translate to accurate predictions of future regime switches. This highlights the limitations of relying solely on univariate data for change point prediction.

In light of previous results, we believe that predictions in real data involve a more complex procedure to identify leading factors that can signal impending regime changes. In the remaining part, we examine a few major market events and their associated leading indicators, demonstrating the potential of a multivariate approach to change point prediction.

- Dot-com Bubble (2000)

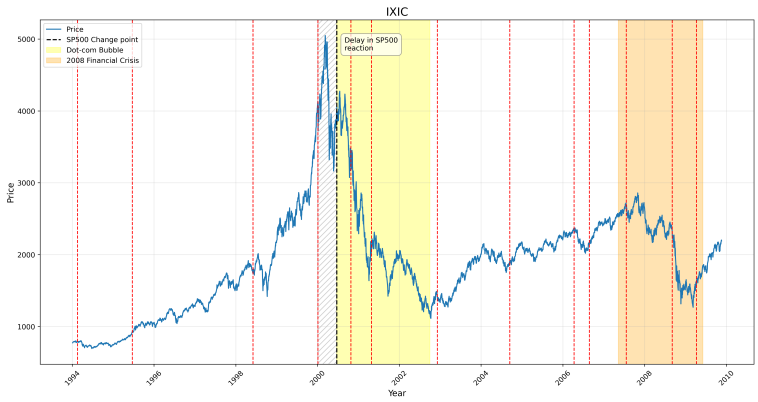


Figure 3.8: NASDAQ composite index from 1994 to 2010

During the late 1990s, the popularisation of the Internet together with low interest rates fuelled a massive influx of capital into technology startups. Many of these

companies had never turned a profit and went out of business after using up all invested capital. Being a technology heavy index, the NASDAQ composite was observed to have preceded the broader market decline, as evidenced by the lag in change points detected between NASDAQ and S&P 500 in Figure 3.8.

- Global Financial Crisis (2008)

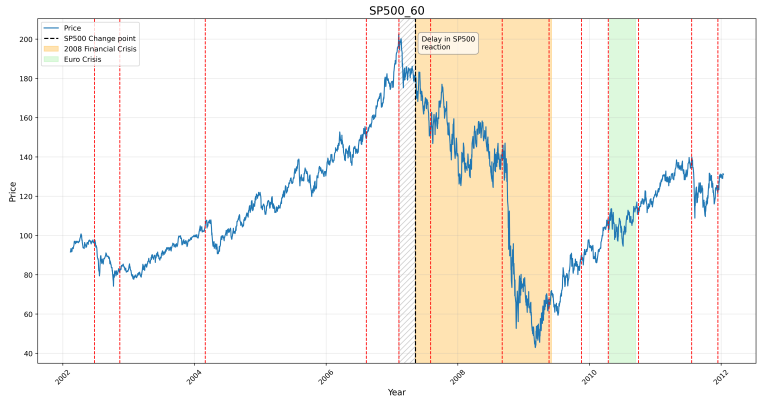


Figure 3.9: S&P 500 real estate sector from 2002 to 2012

The 2008 financial crisis was a multi-faceted disaster caused by inadequate financial regulations and poor mortgage lending standards. Initially set off by the collapse of the U.S. housing market, it quickly evolved into a worldwide economic recession, leading to the loss of trillions of wealth and widespread unemployment. As most of the companies in the S&P500 real estate sector index are real estate investment trusts (REITs), their performances were closely tied to the US housing market, especially in times of distress. Hence, we used it as a proxy for the average US housing price, which has limited availability. As we can see in Figure 3.9, the regime change in the real estate sector leads the main S&P 500 index by almost three months.

- Euro Debt Crisis (2010)

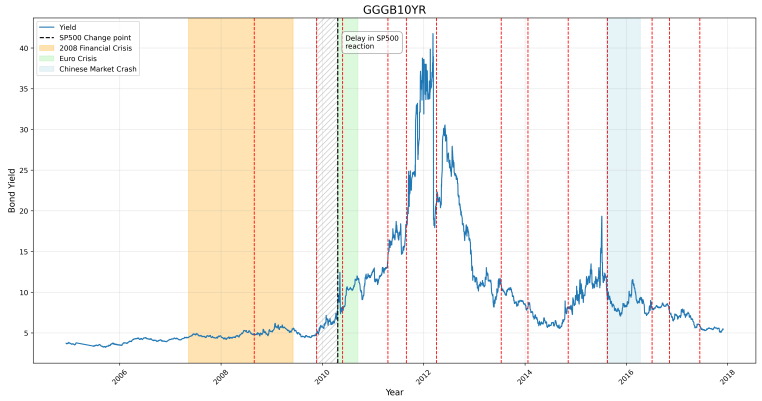


Figure 3.10: Greece 10 year government bond from 2005 to 2018

The accumulation of public debts in many European countries raised concerns over their ability to finance the deficit. In fear of potential default, the Greece government bond experienced a sharp increase in yield during the years of crisis. We observed the structural break of the Greece 10 year bond yield preceded the reaction to the Euro debt crisis in the U.S. market as suggested by the gap in change point detection.

- Chinese Market Crash (2016)

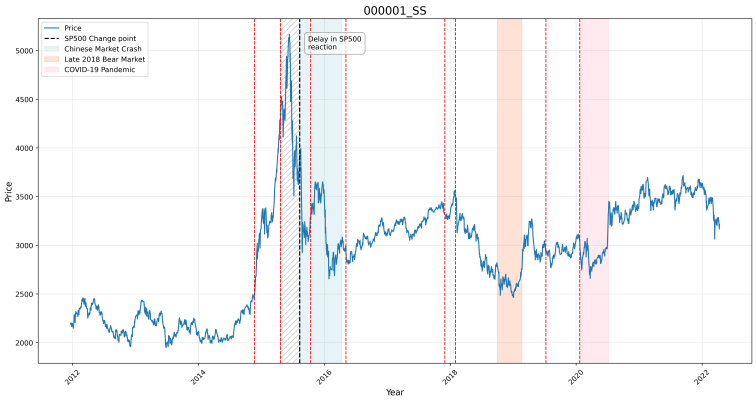


Figure 3.11: SSE composite index from 2012 to 2022

After experiencing a year of rapid growth, the Chinese stock market experienced a significant downturn in 2015, losing more than 30% of its value. The mismatch between excessive speculation and slowing economic growth created a bubble in the Chinese stock market. This was reflected in the steep decline in the Shanghai stock exchange index followed by a period of volatile returns in the S&P 500 index.

It is crucial to acknowledge that our analysis may suffer from storytelling bias. In retrospect, it is relatively easy to identify indicators that seemed to predict past market events. The clear patterns we observed in hindsight may not have been as evident to market participants in real-time, underscoring the complexity of predicting structural breaks in financial markets. As each crisis was preceded by different types of indicators, it is challenging to identify a universal set of leading factors, and a one-size-fits-all approach to change point prediction is likely to be ineffective. However, these findings also point to a promising direction for enhancing our predictive framework. By monitoring a diverse set of economic and financial indicators, we can develop a more comprehensive early warning system for structural breaks, allowing for timely risk management action.

Conclusion

In this study, we conducted an extensive exploration and comparison of various two-sample tests under different scenarios. This provided valuable insights into the strengths and limitations of different methods. In particular, we found that kernel-based approaches such as the Maximum Mean Discrepancy (MMD) and the Kernel Fisher Discriminant ratio (KFDR) tests are more effective in detecting distributional changes. Additionally, we proposed three novel change point detection algorithms based on two-sample tests. Notably, the partitioning method based on rolling statistics is readily adaptable to an online setting and achieves ϵ-real time detection. Lastly, we applied our algorithms to real-world data and verified their effectiveness in identifying known periods of market instability.

We suggest several promising areas for future research. One interesting possibility is to explore the application of MMD and KFDR to multivariate change point detection, extending the current framework to detecting correlation regimes. Incorporating clustering techniques to group similar regimes could enhance our understanding of market dynamics. Currently, our methods identify changes but do not classify the new regimes. The framework outlined in [25], especially the modelling of prior beliefs, could serve as a good starting point for this extension. Finally, exploring ensemble approaches that combine multiple change point detection methods could potentially improve the robustness and accuracy of our results.

Appendix A

Technical Results

A.1 Maximum Mean Discrepancy

Lemma A.1.1. *([17], Lemma 4). If the reproducing kernel $k(\cdot,\cdot)$ of RKHS $\mathcal{H}$ is measurable and $\mathbb{E}_p[\sqrt{k(x,x)}] \leq \infty$, the mean embedding of p exists. Furthermore,*

$$\mathrm{MMD}^2[\mathcal{H}, p, q] = \|\mu_p - \mu_q\|^2_{\mathcal{H}}.$$

Proof.

$$\begin{aligned}
\mathrm{MMD}^2[\mathcal{H}, p, q] &= \left[\sup_{\|f\|_{\mathcal{H}}\leq 1} (\mathbb{E}_p[f(x)] - \mathbb{E}_q[f(y)])\right]^2 \\
&= \left[\sup_{\|f\|_{\mathcal{H}}\leq 1} \langle \mu_p - \mu_q, f\rangle\right]^2 \\
&= \|\mu_p - \mu_q\|^2_{\mathcal{H}}.
\end{aligned}$$

□

Definition A.1.2. (Universality, [17], Section 2.2). Let $(\mathcal{X}, d)$ be a compact metric space. We say $\mathcal{H}$ is a universal RKHS if it is dense in $C(\mathcal{X})$ with respect to the L_∞ norm and its associated kernel $k(\cdot,\cdot)$ is continuous.

Theorem A.1.3. *([17], Theorem 5). Let $\mathcal{F}$ be a unit ball in a universal RKHS $\mathcal{H}$, defined on the compact space $\mathcal{X}$. Then* $\mathrm{MMD}[\mathcal{F}, p, q] = 0$ *if and only if* $p = q$.

Proof. $\implies$ By Lemma A.1.1, $p = q$ clearly implies $\mathrm{MMD}[\mathcal{F}, p, q] = 0$.
$\impliedby$ By universality of $\mathcal{H}$, for any $\epsilon > 0$ and $f \in C(\mathcal{X})$ there exists g such that

$$\|f - g\|_\infty \leq \epsilon.$$

Then

$$|\mathbb{E}_p[f] - \mathbb{E}_q[f]| \leq |\mathbb{E}_p[f] - \mathbb{E}_p[g]| + |\mathbb{E}_p[g] - \mathbb{E}_q[g]| + |\mathbb{E}_q[g] - \mathbb{E}_q[f]| < 2\epsilon.$$

Since

$$|\mathbb{E}_p[f] - \mathbb{E}_q[f]| \leq \mathbb{E}[|f - g|] \leq \|f - g\|_\infty$$

and

$$|\mathbb{E}_p[g] - \mathbb{E}_q[g]| = |\langle g, \mu_p - \mu_q\rangle| = 0.$$

The result follows from Lemma A.1.1. □

A.2 Kernel Fisher Discriminant Ratio

Corollary A.2.1. *(Computation of KFDR, [12], Section 3.2). Given samples* $(X_1, \ldots X_{n_1})$ *and* $(Y_1, \ldots, Y_{n_2})$*. We define matrix* $\mathbf{K}_{ij} := k(X_i, Y_j)$ *for* $i \in 1, \ldots, n_1, j \in 1, \ldots, n_2$ *where* $k(\cdot, \cdot)$ *is the kernel of choice. Also, define vector* $\mathbf{m}$ *where* $\mathbf{m}_i := -n_1^{-1}$ *for* $i = 1, \ldots, n_1$ *and* $:= n_2^{-1}$ *for* $i = 1, \ldots, n_2$*. Lastly, define matrix*

$$\mathbf{N} := \begin{pmatrix} \mathbf{P}_{n_1} & 0 \\ 0 & \mathbf{P}_{n_2} \end{pmatrix}$$

where $\mathbf{P}_\ell = \mathrm{I}_\ell - \ell^{-1}\mathbf{1}_\ell\mathbf{1}_\ell^T$*. Then, the empirical kernel Fisher discriminant ratio is given by*

$$\widehat{\mathrm{KFDR}}[\mathcal{H}, X, Y] = \frac{n_1 n_2}{\gamma n}\left\{\mathbf{m}^T\mathbf{K}\mathbf{m} - n^{-1}\mathbf{m}^T\mathbf{K}\mathbf{N}\left(\gamma\mathrm{I} + n^{-1}\mathbf{N}\mathbf{K}\mathbf{N}\right)^{-1}\mathbf{N}\mathbf{K}\mathbf{m}\right\}$$

where $n = n_1 + n_2$*.*

Theorem A.2.2. *(Limiting distribution under null hypothesis, [12], Theorem 3). Refer to assumptions made in [12] A1, B1, B2. Under the null hypothesis and condition that the sequence* $\{\gamma_n\}$ *follows*

$$\gamma_n + d_2^{-1}(\Sigma_W, \gamma_n)d_1(\Sigma_W), \gamma_n)\gamma_n^{-1}n^{-\frac{1}{2}} \to 0,$$

$$\widehat{T}_n(\gamma_n) \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1).$$

Appendix B

Further results

B.1 Comparison of power

- Type I error under normal distribution with mean and variance 1

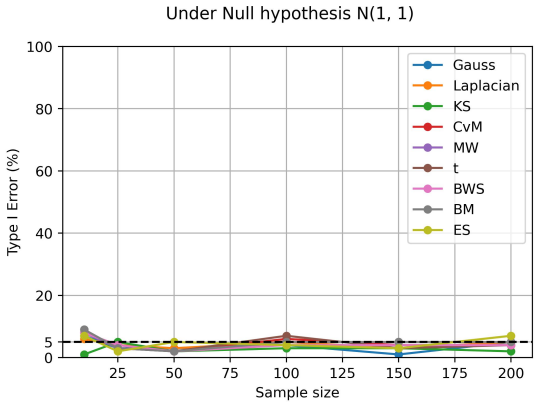
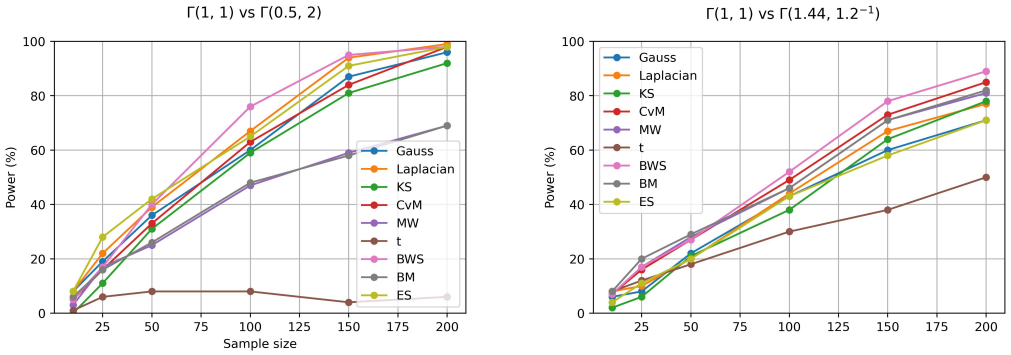


Figure B.1: Type I error (%) against sample size at level $\alpha = 5\%$

- Mean and variance change under Gamma distributions



(a) Mean change

(b) Variance change

Figure B.2: Power (%) against sample size at level $\alpha = 5\%$

- Effect of sample ratio on power under change in mean

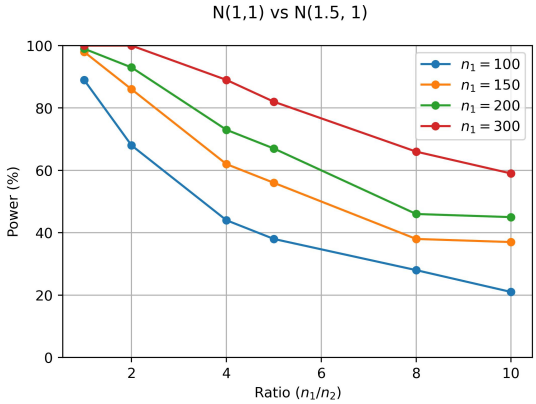


Figure B.3: Power (%) against ratio between sample sizes of two sets of independent observations at level $\alpha = 5\%$

B.2 Kernel Fisher Discriminant Ratio

- Type I error under normal distribution with mean 0 and variance 1

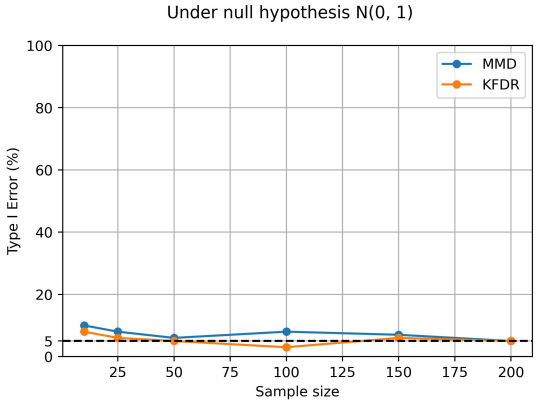


Figure B.4: Type I error (%) against sample size at level $\alpha = 5\%$

- Comparison of power between MMD and KFDR two-sample tests

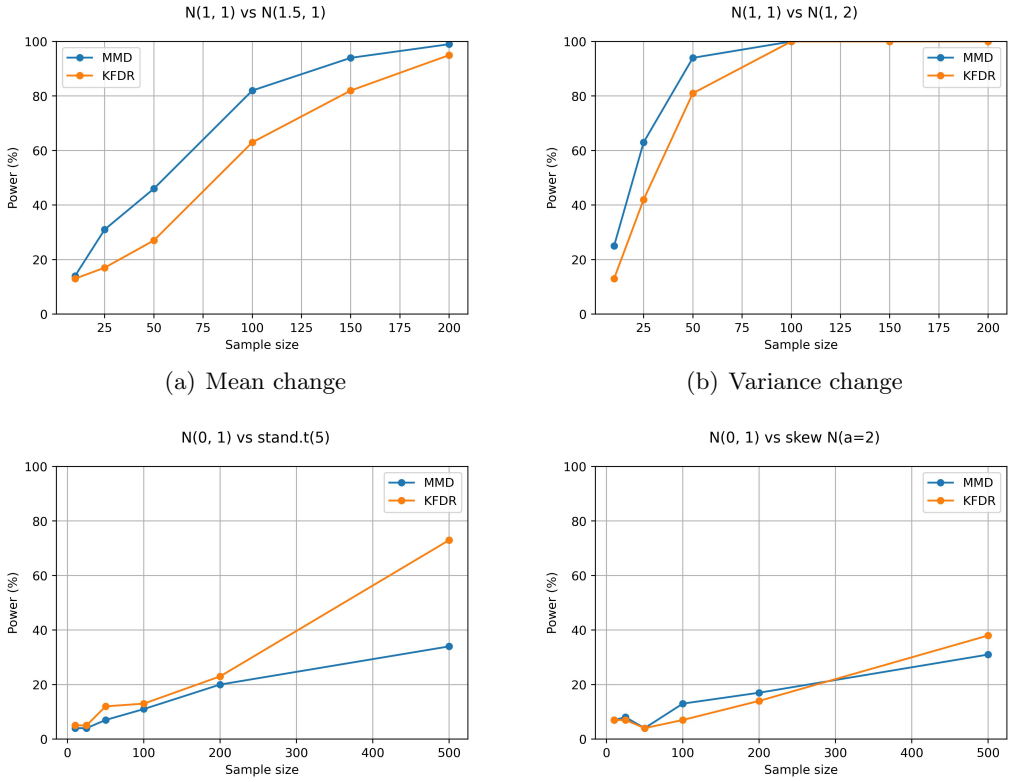
  Results provided in Figure B.5.

B.3 Algorithm

We noticed that the batch size maximising the usage of my GPU memory is 20. Any size larger than 20 resulted in significant deterioration of speed performance. The upper and lower bounds are standard deviations estimated with 10 runs. Results provided in Figure B.6.

B.4 Asymmetric weighting scheme

We considered a window size of 400 data points on either side of the change points and applied a weighting scheme, where samples before the change points were reweighed by 0.1, and samples after the change points were reweighed by 1.5. Table B.1 below presents the true negative rates restricted to the preceding and succeeding period with and without the aforementioned weighting scheme.



(a) Mean change

(b) Variance change

(c) Kurtosis change

(d) Skewness change

Figure B.5: Power (%) against sample size at level $\alpha = 5\%$

	GBM	MJD	SGT
Accuracy preceding (unweighted)	0.68	0.72	0.66
Accuracy preceding (weighted)	0.63	0.68	0.63
Accuracy after (unweighted)	0.53	0.65	0.55
Accuracy after (weighted)	0.59	0.69	0.57

Table B.1: True negative rates under various dynamics

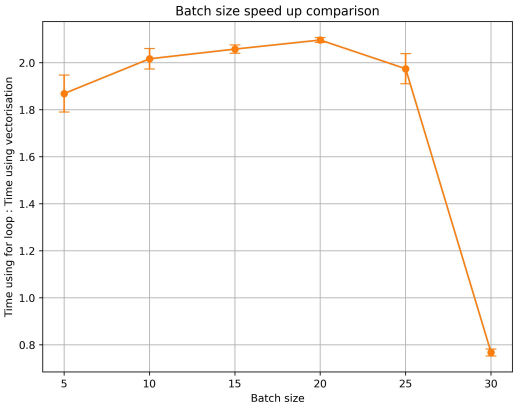


Figure B.6: Improvement ratio for varying batch size

Bibliography

[1] T. W. Anderson, *On the distribution of the two-sample cramer-von mises criterion*, The Annals of Mathematical Statistics, (1962), pp. 1148–1159.

[2] A. Ang and A. Timmermann, *Regime changes and financial markets*, Annu. Rev. Financ. Econ., 4 (2012), pp. 313–337.

[3] N. Aronszajn, *Theory of reproducing kernels*, Transactions of the American mathematical society, 68 (1950), pp. 337–404.

[4] D. A. Bodenham and Y. Kawahara, *eummd: efficiently computing the mmd two-sample test statistic for univariate data*, Statistics and Computing, 33 (2023), p. 110.

[5] E. Brunner and U. Munzel, *The nonparametric behrens-fisher problem: asymptotic theory and a small-sample approximation*, Biometrical Journal: Journal of Mathematical Methods in Biosciences, 42 (2000), pp. 17–25.

[6] T. Cass and C. Salvi, *Lecture notes on rough paths and applications to machine learning*, arXiv preprint arXiv:2404.06583, (2024).

[7] A. Celisse, G. Marot, M. Pierre-Jean, and G. Rigaill, *New efficient algorithms for multiple change-point detection with reproducing kernels*, Computational Statistics & Data Analysis, 128 (2018), pp. 200–220.

[8] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, *Smote: synthetic minority over-sampling technique*, Journal of artificial intelligence research, 16 (2002), pp. 321–357.

[9] I. Chevyrev and A. Kormilitzin, *A primer on the signature method in machine learning*, arXiv preprint arXiv:1603.03788, (2016).

[10] R. M. Dudley, *Convergence of Laws and Central Limit Theorems*, Cambridge University Press,, Cambridge :, [new ed.] ed., 2010.

[11] T. Epps and K. J. Singleton, *An omnibus test for the two-sample problem using the empirical characteristic function*, Journal of Statistical Computation and Simulation, 26 (1986), pp. 177–203.

[12] M. Eric, F. Bach, and Z. Harchaoui, *Testing for homogeneity with kernel fisher discriminant analysis*, Advances in Neural Information Processing Systems, 20 (2007).

[13] K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf, *Kernel measures of conditional dependence*, Advances in neural information processing systems, 20 (2007).

[14] K. Garg, J. Yu, T. Behrouzi, S. Tonekaboni, and A. Goldenberg, *Dynamic interpretable change point detection*, arXiv preprint arXiv:2211.03991, (2022).

[15] D. Garreau, W. Jitkrittum, and M. Kanagawa, *Large sample analysis of the median heuristic*, arXiv preprint arXiv:1707.07269, (2017).

[16] S. J. Goerg and J. Kaiser, *Nonparametric testing of distributions—the epps–singleton two-sample test using the empirical characteristic function*, The Stata Journal, 9 (2009), pp. 454–465.

[17] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, *A kernel two-sample test*, The Journal of Machine Learning Research, 13 (2012), pp. 723–773.

[18] A. Gretton, D. Sejdinovic, H. Strathmann, S. Balakrishnan, M. Pontil, K. Fukumizu, and B. K. Sriperumbudur, *Optimal kernel choice for large-scale two-sample tests*, Advances in neural information processing systems, 25 (2012).

[19] J. D. Hamilton, *A new approach to the economic analysis of nonstationary time series and the business cycle*, Econometrica: Journal of the econometric society, (1989), pp. 357–384.

[20] Z. Harchaoui, F. Bach, O. Cappe, and E. Moulines, *Kernel-based methods for hypothesis testing: A unified view*, IEEE Signal Processing Magazine, 30 (2013), pp. 87–97.

[21] Z. Harchaoui and O. Cappé, *Retrospective mutiple change-point estimation with kernels*, in 2007 IEEE/SP 14th Workshop on Statistical Signal Processing, IEEE, 2007, pp. 768–772.

[22] Z. Harchaoui, E. Moulines, and F. Bach, *Kernel change-point analysis*, Advances in neural information processing systems, 21 (2008).

[23] W. Hoeffding, *Probability inequalities for sums of bounded random variables*, The collected works of Wassily Hoeffding, (1994), pp. 409–426.

[24] B. Horvath, Z. Issa, and A. Muguruza, *Clustering market regimes using the wasserstein distance*, arXiv preprint arXiv:2110.11848, (2021).

[25] Z. Issa and B. Horvath, *Non-parametric online market regime detection and regime clustering for multidimensional and path-dependent data structures*, arXiv preprint arXiv:2306.15835, (2023).

[26] M. Kritzman, S. Page, and D. Turkington, *Regime shifts: Implications for dynamic strategies (corrected)*, Financial Analysts Journal, 68 (2012), pp. 22–39.

[27] R. Lopes, I. Reid, and P. Hobson, *The two-dimensional kolmogorov-smirnov test*, XI International Workshop on Advanced Computing and Analysis Techniques in Physics Research, (2007).

[28] A. Lung-Yut-Fong, C. Lévy-Leduc, and O. Cappé, *Homogeneity and change-point detection tests for multivariate data using rank statistics*, arXiv preprint arXiv:1107.1971, (2011).

[29] H. B. Mann and D. R. Whitney, *On a test of whether one of two random variables is stochastically larger than the other*, The annals of mathematical statistics, (1947), pp. 50–60.

[30] J. Mc Greevy, A. Muguruza, Z. Issa, C. Salvi, J. Chan, and Z. Zuric, *Detecting multivariate market regimes via clustering algorithms*, Available at SSRN 4758243, (2024).

[31] M. Neuhäuser, *Exact tests based on the baumgartner-weiß-schindler statistic—a survey*, Statistical Papers, 46 (2005), pp. 1–29.

[32] Y. S. Niu, N. Hao, and H. Zhang, *Multiple change-point detection: a selective overview*, Statistical Science, (2016), pp. 611–623.

[33] O. H. M. Padilla, Y. Yu, D. Wang, and A. Rinaldo, *Optimal nonparametric change point detection and localization*, arXiv preprint arXiv:1905.10019, (2019).

[34] A. Schrab, I. Kim, M. Albert, B. Laurent, B. Guedj, and A. Gretton, *Mmd aggregated two-sample test*, Journal of Machine Learning Research, 24 (2023), pp. 1–81.

[35] P. Theodossiou, *Financial data and the skewed generalized t distribution*, Management science, 44 (1998), pp. 1650–1661.

[36] C. Truong, L. Oudre, and N. Vayatis, *ruptures: change point detection in python*, arXiv preprint arXiv:1801.00826, (2018).

[37] ——, *Selective review of offline change point detection methods*, Signal Processing, 167 (2020), p. 107299.

[38] Wikipedia contributors, *Reproducing kernel hilbert space — Wikipedia, the free encyclopedia*. `https://en.wikipedia.org/w/index.php?title=Reproducing_kernel_Hilbert_space&oldid=1243483630`, 2024. [Online; accessed 3-September-2024].

[39] ——, *Student's t-test — Wikipedia, the free encyclopedia*. `https://en.wikipedia.org/w/index.php?title=Student%27s_t-test&oldid=1242239938`, 2024. [Online; accessed 2-September-2024].

[40] C. Zou, G. Yin, L. Feng, and Z. Wang, *Nonparametric maximum likelihood approach to multiple change-point problems*, The Annals of Statistics, 42 (2014).