Clustering volatility regimes for dynamic trading strategies

Gilad Francis[a,*], Nick James[a,*], Max Menzies[b,*], Arjun Prakash[a,*]

[a] *School of Mathematics and Statistics, University of Sydney, NSW, Australia*
[b] *Yau Mathematical Sciences Center, Tsinghua University, Beijing, China*

Abstract

We develop a new method to find the number of volatility regimes in a non-stationary financial time series. We use change point detection to partition a time series into locally stationary segments, then estimate the distributions of each piece. The distributions are clustered into a learned number of discrete volatility regimes via an optimisation routine. Using this method, we investigate and determine a clustering structure for indices, large cap equities and exchange-traded funds. Finally, we create and validate a dynamic portfolio allocation strategy that learns the optimal match between the current distribution of a time series with its past regimes, thereby making online risk-avoidance decisions in the present.

Keywords: Volatility modelling, Spectral clustering, Non-parametric method, Dynamic portfolio strategy, Change point detection

1. Introduction

Modelling the volatility of a financial time series is an important task for traders and economists. Financial markets are not only important in their own right, but also have immense flow on effects on the rest of society, as seen during the global financial crisis, US-China trade war or current COVID-19 pandemic. Volatility may be modelled from an individual stock level to an index level; the latter can represent the uncertainty of an entire sector or economy. It has been surmised that financial time series exhibit regime switching patterns, switching between periods of heightened volatility and ease [1, 2, 3].

Statistical methods for volatility modelling have long been popular in the literature [4]. Long standing parametric methods such as ARCH and GARCH [5, 6] model the volatility of individual stocks, designing models in order to obey assumptions such as *stylized facts* [7], and appropriately choosing parameters to best fit past data. This allows traders to model future returns, assuming these assumptions continue to hold. *Regime switching models* [1, 8] have been

*Equal contribution

developed to model the patterns of volatility switching; these are also generally parametric, building on ARCH and GARCH, and must *a priori* estimate the number of regimes. In this paper, we introduce a new *non-parametric* method to analyse the number and switching behaviour of volatility regimes, making no assumptions about the data or underlying context.

We begin by decomposing our data into locally stationary segments, via *change point detection*. Developed by Hawkins et al. [9, 10], change point algorithms seek to determine breaks in a time series at which the stochastic properties of the underlying random variables change, and have become instrumental in time series analysis. Specifically, in order to analyse volatility, we use the *Mood test for variance*, Section 4 of [11].

Having determined structural breaks, we associate to each segment of the partitioned time series a distribution. Then, we use the Wasserstein metric to compute the distances between these distributions and use spectral clustering to allocate these segments into specific classes of volatility regimes. The precise number of regimes is carefully chosen. Thus, our method can determine the number of regimes to use in any candidate regime switching model in advance. We then draw on our findings and use additional learning procedures to design a dynamic trading strategy. We show that it provides superior risk-adjusted returns to the S&P 500 index in various market conditions. This improves on existing strategies in two ways detailed below. Our contributions are as follows:

1. We introduce a non-parametric method for detection and classification of volatility regimes, including the number of regimes.

2. We include validation scores for our methodology, and demonstrate good results for synthetic data and real data across a variety of asset classes that match well with known periods of higher volatility.

3. We develop a new dynamic trading strategy that is able to identify volatile time periods and allocate capital in real-time. By learning the past volatility structure of the S&P 500, we determine whether the present time period is volatile based on the minimal distance to other past distributions.

4. This improves on existing methods in two ways. First, it is more reliable than simply switching at a detected change point [12], as a change point may not indicate a change in volatility regime. Secondly, we optimise the time period of how long to look back; a change point algorithm also has a detection delay, but it cannot be controlled.

In Section 2, we outline all steps of our methodology in technical detail. In Section 3, we validate our methodology on synthetic data, and then show a reasonable clustering structure can be determined for major indices, stocks, and popular ETFs. In Section 4, we develop our dynamic portfolio allocation trading strategy, incorporating our insights and additional learning procedures. Section 5 concludes the body of the paper. In Appendix A and Appendix B respectively, we provide a description of the change point algorithm used and provide additional figures from our experiments.

2. Mathematical model

In mathematical statistics, a *time series* (X_t) is a sequence of random variables - measurable functions from a probability space to the real numbers - indexed by time. In finance, one generally conflates the random variable with the observed data point at each point in time. As such, a *financial time series* is a sequence of price data. In this paper, we will examine the time series of adjusted closing prices $(p_t)_{t\geq 0}$ at time t, and the log returns $(R_t)_{t\geq 1} = \log(\frac{p_t}{p_{t-1}})$.

In subsequent sections, we describe our method in detail. We begin by assuming our non-stationary time series are generated from Dahlhaus locally stationary processes [13] and proceed to partition the time series into stationary segments; specifically, we detect changes in the volatility of a time series via the Mood test change point method. We then estimate the distribution of each segment via kernel density estimation, and use the Wasserstein metric to quantify distance between these distributions. We determine an allocation into an appropriate number of clusters by an optimisation routine that combines spectral clustering and silhouette scoring. Thus, we classify our segments of volatility into discrete classes in a non-parametric way. We record the number of clusters and their structure, together with the silhouette score as a means of validating the allocation into these volatility regimes.

The precise method, applicable to volatility clustering, that we describe below, is not exhaustive. As long as there is consistency between the regime characteristic of interest, the change point algorithm (and its test statistic if applicable), and the distance metric between distributions, the method below could easily be reworked for detection and classification of regimes of alternative characteristics.

2.1. Partition of the time series

Given time series price data, begin by forming the log return time series $(R_t)_{t=1,...,T}$ over a particular time interval. It is generally appropriate to assume the log returns are independent random variables, but not appropriate to assume they have any particular distribution. With this in mind, we apply the non-parametric *Mood test*, performed in the CPM package of Ross [14], to detect changes in the volatility of a time series. Although this is commonly known as a median test, it is also appropriate for detecting change in the variance between two distributions, as described in Section 4 of [11]. More details on the change point framework and implementation can be found in Appendix A. This yields a collection of change points $\tau_1, ..., \tau_{m-1}$. For notational convenience, set $\tau_0 = 1, \tau_m = T$. The stationary segments according to this partition are then

$$(R_t)_{t\in[\tau_{j-1},\tau_j]}, j = 1, 2, ..., m$$

This yields m stationary segments. Now let $(Y^{(j)})$ be the restricted time series whose entries are taken from the time interval $[\tau_{j-1}, \tau_j]$. That is, $(Y_t^{(j)})$ consists of the values R_t where t ranges from τ_{j-1} to τ_j. Each $(Y^{(j)})$ has been determined by the algorithm to be sampled from a consistent distribution.

2.2. Kernel density estimation

Next, for each stationary segment $(Y^{(j)})$, we perform *kernel density estimation* to estimate the probability density function of the underlying distribution. In general, given data points $(x_1, x_2, ..., x_n)$ drawn from some arbitrary data generating process, the KDE is given by:

$$f_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

K is a kernel function, h is a smoothing parameter. We use a Gaussian kernel for K, [15] and the Silverman rule of thumb [16] to choose h.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$
$$\bar{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$
$$\bar{h} = \left(\frac{4\bar{\sigma}^5}{3n}\right)$$

With this procedure, we associate to each restricted time series $(Y^{(j)})$ a kernel density function $f^{(j)}, j = 1, ..., m$.

2.3. Wasserstein distance

Next, we compute the Wasserstein distance between these kernel density functions $f^{(j)}$. The Wasserstein metric, also known as the earth mover's distance, is the minimal work to move the mass of one probability distribution into another. Given probability measures μ, ν on Euclidean space $\mathbb{R}^d$, define

$$W_p(\mu, \nu) = \inf_{\gamma}\left(\int_{\mathbb{R}^d \times \mathbb{R}^d} ||x - y||^p d\gamma\right)^{\frac{1}{p}}.$$

This infimum is taken over all joint probability measures γ on $\mathbb{R}^d \times \mathbb{R}^d$ with marginal probability measures μ and ν. In our case, $d = 1$. To each kernel density estimate function $f^{(j)}$, we form the associated Radon-Nikodym measure [17] $\mu_j = f^{(j)}(x)dx$, where dx is Lebesgue measure. This allows us to form a $m \times m$ distance matrix of Wasserstein distances.

$$D_{ij} = W_p(\mu_i, \mu_j) = W_p(f^{(i)}dx, f^{(j)}dx)$$

Henceforth, set $p = 1$. Concretely, $W_p(f(x)dx, g(x)dx)$ may be computed [18] as

$$\int_{\mathbb{R}} |F - G|dx$$

where F, G are the cumulative density functions associated to probability density functions f and g respectively. Thus, we produce an $m \times m$ distance matrix D between the m distributions of each locally stationary segment of the log return time series. The Wasserstein metric is continuous with respect to small perturbations in the probability density functions, so small changes to the kernel density estimates through choices of h and K above affect the distances only slightly.

2.4. Spectral clustering

To the distance matrix D we associate an *affinity matrix* A by

$$A_{i,j} = \exp\left(\frac{-D_{i,j}^2}{2\sigma^2}\right)$$

where σ is a parameter to be chosen. One then forms the *Laplacian* L and *normalized Laplacian* L_{sym} following [19]. First, form the diagonal degree matrix given by $\text{Deg}_{ii} = \sum_j A_{ij}$. Then form

$$L = \text{Deg} - A$$
$$L_{\text{sym}} = \text{Deg}^{-1/2} L \text{Deg}^{-1/2}$$

Note L, L_{sym} are $m \times m$ symmetric matrices, and hence are diagonalizeable with all real eigenvalues. By the definition of L and the normalization L_{sym}, all their eigenvalues are non-negative, $0 = \lambda_1 \leq ... \leq \lambda_m$. Spectral clustering proceeds as follows. For some fixed choice of k, compute the normalized eigenvectors $u_1, ..., u_k$ corresponding to the k smallest eigenvalues of L_{sym}. Form the matrix $U \in \mathbb{R}^{m \times k}$ whose columns are $u_1, ..., u_k$. Let $v_i \in \mathbb{R}^k$ be the rows of U, $i = 1, ..., m$. Cluster these rows into clusters $C_1, ..., C_k$ according to k-means. Finally, output clusters $A_l = \{i : v_i \in C_l\}, l = 1, ..., k$ to assign the original m elements, in this case segment KDEs, into the corresponding clusters.

2.5. Choice of k and silhouette scoring

Spectral clustering has a uniquely determined output (in the absence of degeneracy) given a fixed k, but the choice of optimal k is a problem with no definitive answer. We introduce the concept of *silhouette scoring* [20]. Suppose m data nodes indexed $1, 2, ..., m$ have been sorted into k clusters $C_l, l = 1, ..., k$. Following the notation of the previous sections, let D_{ij} be the distances between these nodes. For a node i in cluster C, define an internal cluster distance by

$$a(i) = \frac{1}{|C| - 1} \sum_{j \in C \setminus \{i\}} D_{ij}$$

Next, define $b(i)$ as the minimal dissimilarity between node $i \in C$ and any different cluster C',

$$b(i) = \min_{C' \neq C} \frac{1}{|C'|} \sum_{j \in C'} D_{ij}$$

The silhouette score for the point i is defined as:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

Finally, the overall silhouette score of the clustering is simply the average $s = \frac{1}{m}\sum_{i=1}^{m} s(i)$. Note each individual $s(i)$ ranges from -1 to 1. The closer it is to 1, the better matched the node i to its constituent cluster C. $s(i) = 0$ is poor and $s(i) = -1$ is abysmal, so the value should be as close to 1 as possible. Therefore, the final s is an overall score for the quality of the clustering. Table 1 records the interpretation of these values, as in [21].

In order to select k, we combine two methods. We begin with a standard choice of $\sigma = 1$. With this parameter set, we use the *elbow method* [19] within our spectral clustering implementation in order to select a first choice k_0. We then identify the respective clusters identified relative to this value k_0. Then, we use this initial estimate as the starting point for an optimisation routine. We vary k and σ simultaneously, at each point recording the respective cluster outputs, and calculating the total silhouette score for that clustering. We vary our parameters in the range $2 \leq k \leq 5, 0.1 \leq \sigma \leq 10$, in order to optimise the silhouette score among the determined cluster outputs.

Our entire method concludes with outputting the clusters as well as the silhouette score as a means of validation. Therefore, we have partitioned a time series into segments according to changes in their volatility, clustered those segments that are similar in distribution with respect to the Wasserstein metric, and included a validation metric for the quality of the clustering.

Silhouette Score	Interpretation
0.71 - 1.00	A strong structure has been found
0.51 - 0.70	A reasonable structure has been found
0.26 - 0.50	A weak and possibly artificial structure has been found
≤ 0.25	No substantial structure has been found

Table 1: Silhouette score interpretation

3. Results

3.1. Synthetic data

In this section, we validate our method on a synthetic time series. We generate this time series, with artificially pronounced breaks in volatility, by concatenating different segments, each randomly drawn from two data generating processes and randomly chosen between 150 and 200 in length:

$$X_1 \sim \mathcal{N}(0 + \epsilon_1, 0.2 + \epsilon_2) \text{ or } X_2 \sim \mathcal{N}(0 + \epsilon_3, 0.01 + \epsilon_4)$$

ϵ_i are added random noise to ensure none of the data generating processes are identical. This time series, together with the change point partition described in Section 2.1, is displayed in Figure 1a. In this case, the delay between change point and detection time, described in detail in Appendix A, is not visible.

Subsequently, we form the kernel density estimate functions, compute the Wasserstein distance, and perform the clustering of the resulting distributions. Figure 1b shows the KDEs on one plot; they have been clustered into two clusters and coloured accordingly. Figure 1c shows the final clustering of the segments of the synthetic time series. Note this whole procedure correctly identifies the change in variance, as well as the existence of two regimes (clusters) of volatility. The final silhouette score in this synthetic example is an excellent 0.91.

3.2. S&P 500

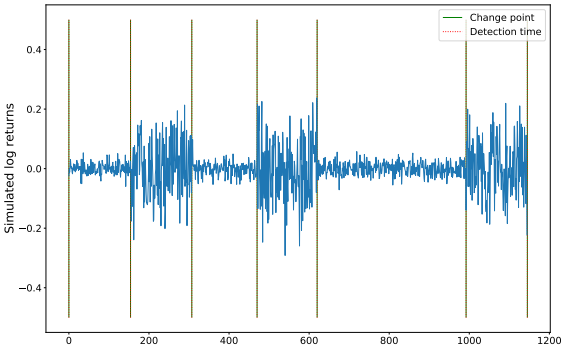
In this section, we apply our method to the S&P 500, and analyse the volatility clustering in detail. We draw adjusted closing price data from Yahoo! Finance, `https://finance.yahoo.com`, from 1 October 2009 to 1 October 2019, and immediately calculate the log returns. We begin by forming the change point segmentation of the time series as in the previous section, and clustering the KDEs, displayed in Figures 2a and 2b respectively. Only two clusters are found, one of lower volatility, one of greater volatility. The silhouette score for this clustering is 0.63, indicating a reasonable structure has been found.

The blue periods of higher volatility correspond to the 2010 flash crash, the European sovereign debt crisis of August 2011, the 2015 August flash crash and the US/China trade war in 2018. In Figure 2b, these correspond to the blue kernel density estimate functions. Note these KDEs are more spread out, indicating that their corresponding distributions have much higher variance than the other cluster. Also, note that a change point is detected between the fifth and sixth segment of Figure 2c, and yet there was no regime change in volatility at this time. This can occur when the distributions are different, but not different enough to warrant an entire regime change. Understanding and being able to predict the volatility of the S&P 500 is the basis of our dynamic trading strategy, which will be described in more detail in Section 4.

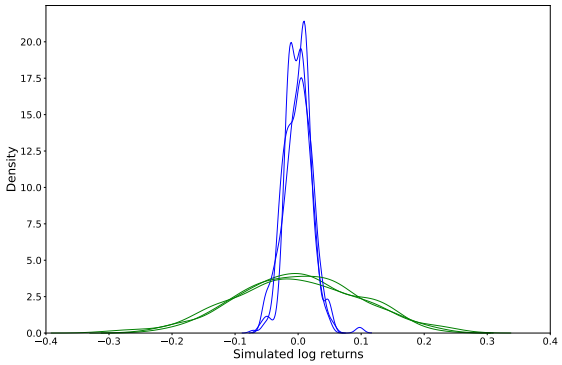
3.3. Empirical results on various asset classes

In this section, we outline the results of our methodology across various asset classes: stocks, currencies, ETFs, and indices. Once again we pull the adjusted closing price data from Yahoo! Finance from 1 October 2009 to 1 October 2019 and calculate the log returns. For each time series, the main result is the number of segments and clusters. This provides the number of discretised volatility regimes. The main evaluation metric is the silhouette score, to two significant figures. We also include the cluster sizes for completeness.

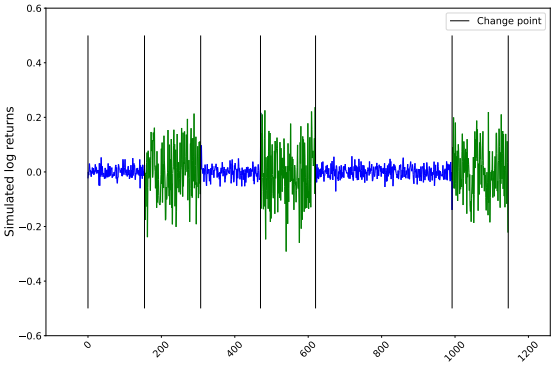
We display these results in Tables 2, 3, 4, 5. In Table 2, the MSCI index is a weighted composite of the 1655 most valuable companies from around the world. Table 3 displays results for large firms: MSFT (Microsoft), APPL (Apple), AMZN (Amazon), GOOG (Alphabet), BRK-A (Berkshire Hathaway



(a) Synthetic time series change points

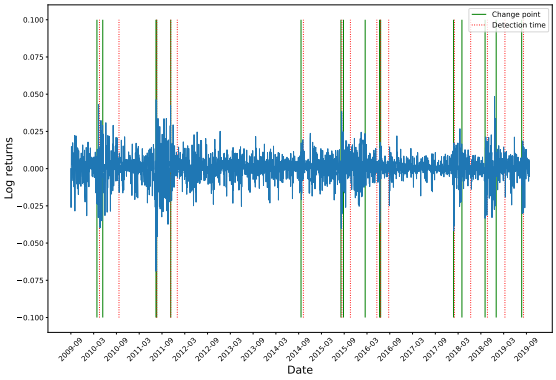


(b) Synthetic distributions forming two clusters

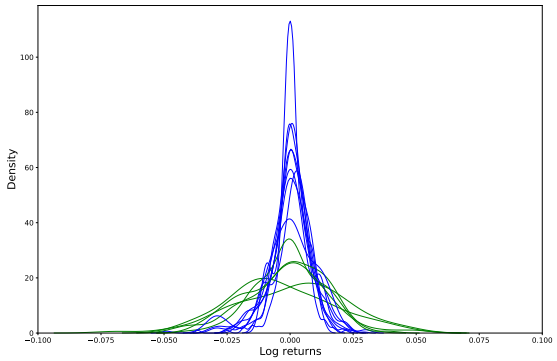


(c) Two determined volatility regimes

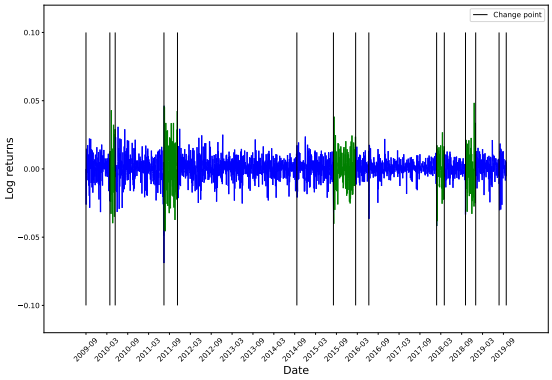
Figure 1: Synthetic data experiment



(a) S&P 500 change points and detection times



(b) S&P 500 clustered distributions



(c) S&P 500 two determined volatility regimes

Figure 2: S&P 500 time series data

Class A). Table 4 displays results for ETFs: RYT (Invesco S&P 500 Equal Weight Technology ETF), GLD (SPDR Gold Shares), XLF (Financial Select Sector SPDR Fund), IJS (iShares SP Small-Cap 600), DGRW (WisdomTree U.S. Quality Dividend Growth Fund). In a testament to the reliability of the clustering algorithm, the choice of σ did not affect any of the cluster outputs. Since the silhouette score is only a function of the cluster outputs, it was not affected by the parameter σ either. Although our methodology described in Section 2 can output the final value of σ, we have omitted it from the tables below. For further reference, all plots of clustered distributions and the time series partitioned into volatility regimes can be found in Appendix B.

3.4. Discussion

According to [21], any silhouette score above 0.5 indicates that a reasonable clustering structure has been found. Hence, the results are promising for indices, large equities and ETFs, which have average silhouette scores of 0.62, 0.63 and 0.60 respectively, with each individual time series among them scoring over 0.5. The results for the currency pairs are slightly weaker, with an average silhouette score of 0.49, but still, three out of five tested pairs have scores of at least 0.5. Remarkably, all time series examined have only two volatility regimes. As an aside, this is by no means inevitable; indeed, by selecting contrived values for the parameter σ such as 0.001, three volatility regimes could be identified. As in Section 3.2, note a change point does not necessarily indicate a change in volatility regime.

Though the count of distributions differs within each cluster, it is still possible to find similarities between related assets. For example, the S&P 500 and the Dow Jones both have volatile periods around March 2010, April 2011, and late 2018. In fact, all five of the listed firms registered a volatile period associated with the US/China trade war of late 2018. In contrast, ETFs do not share many volatile periods, as they are composites of different asset classes.

Regime switching models usually *a priori* assume the number of regimes, for which they are often criticised. When misspecified, they may perform badly, such as modelling three piecewise autoregressive processes with a 2-regime switching model. Interestingly, our model, which estimates the number of data generating processes flexibly, identifies a manageable number of regimes in most scenarios - usually 2. This finding suggests that regime switching models may have their place in statistical modelling for financial time series, if there is a thoughtful way of estimating the number of regimes based on the data. These findings support the work of [1, 3].

Note that our methodology may be combined with other regime switching modelling methods. As our methodology suitably determines the number of volatility regimes, this number can then be used in any other regime switching method, which often require the number of clusters to begin with. As a further application, we show in the next section how these results can be used to make decisions about asset allocation in a dynamic trading strategy.

Ticker	Sil	No. Segments	No. Clusters	Cluster Sizes
S&P 500	0.63	14	2	9,5
Dow Jones	0.66	17	2	12,5
Nikkei 225	0.58	15	2	9,6
FTSE 100	0.66	15	2	10,5
MSCI	0.71	9	2	8,1
Average	**0.62**	**14**	**2**	n/a

Table 2: Major indices

Ticker	Sil	No. Segments	No. Clusters	Cluster Sizes
MSFT	0.63	11	2	6,5
APPL	0.64	12	2	10,2
AMZN	0.68	10	2	7,3
GOOG	0.52	9	2	5,4
BRK-A	0.67	11	2	8,3
Average	**0.63**	**10.4**	**2**	n/a

Table 3: Large firms by market capitalisation

Ticker	Sil	No. Segments	No. Clusters	Cluster Sizes
RYT	0.51	14	2	9,5
GLD	0.52	13	2	9,4
XLF	0.57	16	2	14,2
IJS	0.68	15	2	13,2
DGRW	0.69	11	2	7,4
Average	**0.60**	**13.2**	**2**	n/a

Table 4: Popular ETFs

Ticker	Sil	No. Segments	No. Clusters	Cluster Size
USD/JPY	0.46	12	2	7,5
AUD/USD	0.41	9	2	5,4
EUR/USD	0.55	10	2	7,3
GBP/USD	0.45	9	2	6,3
NZD/AUD	0.50	7	2	5,2
Average	**0.49**	**9.2**	**2**	n/a

Table 5: Currency pairs

4. Application of results: trading strategy

In recent times, passive investing has gathered more asset inflows than active investment management. In particular, index funds and ETFs that track major indices such as the S&P 500 are a popular way of attaining broad market exposure for investors. We apply our analysis of the cluster structure of the S&P 500 index volatility to determine a dynamic trading strategy that can simultaneously benefit from the index's appreciation while minimising risk. In Section 3.2, we determined that the S&P 500 has two distinct volatility regimes, captured in two distinct clusters of volatility periods. Our contrived trading strategy is to buy and hold SPY, a tracker of the S&P 500, in low volatility periods, and then flee to the safe haven of GLD, a gold bullion tracker, in high volatility periods. Note that if our trading strategy were applied among a collection of less efficient assets, such as the index's underlying equity constituents, the trading strategy may attain greater expected returns and higher risk-adjusted return ratios. We improve on the previous work of [12], who uses a live implementation of the rank test to move away from the S&P 500. This method has two drawbacks: first, as noted in Section 3, a change point does not necessarily indicate a change in regime; secondly, their method has an unpredictable delay in registering the change point, as discussed in Appendix A.

Instead, we implement a dynamic procedure with a 4-year sliding window. Model parameters are learned within the prior window, and then applied to the proceeding four years of data. Suppose our algorithm begins with years $0:4$. First, analyse the S&P 500 over the prior 4-year period, years $-4:0$. Determine the cluster structure of the distribution segments of the S&P 500 over this prior period. To make investment decisions in the current period of $0:4$ years, we try to match the present distribution with the most similar distribution in the prior window. Specifically, we examine the present local distribution of the last n days, where n is a learned parameter, and determine the minimal distance between the local distribution and the kernel density estimate distributions of the prior 4-year period. If this closest point lies in the most volatile class of past distributions, characterised by widest kernel density estimate functions, we determine that the local distribution is volatile, and allocate all capital toward gold. This method works even if greater than 2 volatility clusters are found during the previous window.

The parameter n is optimised relative to the $-4:0$ year window. Specifically, having determined the cluster structure, n is chosen to optimise the *Sharpe ratio*, a well-established measure of risk-adjusted returns, when testing over that window. We optimise n over a range $10 \leq n \leq 30$, that is, 2 to 6 trading weeks. Thus, n is learned in this prior window and then used in the algorithm in the subsequent window. The window is then successively slid forward four years, and the process repeats. That is, model parameters estimated on years $0:4$ are used to forecast in years $4:8$, and so on. This 4-year period is chosen as the literature suggests that equity markets follow four year cycles, associated with the cyclicality of Kitchin cycles [22] and the US presidential election [23]. This adaptive sliding window technique allows us to convincingly validate the

long-run performance of our trading strategy.

We analyse the strategy's performance in a period from immediately prior to the global financial crisis (GFC), up to the present day. Accordingly, our initial backtest period of -4 : 0 is 2004-2008, while our first period of trading, years 0 : 4, is 2008-2012. We compare the performance of our dynamic trading strategy with three other strategies: holding SPY, holding GLD, and a baseline strategy holding an equal split between the two. We use six common validation metrics to evaluate and compare our trading strategy.

1. Annualised return (AR): the total return a strategy yields relative to the time the strategy has been in place.

2. The overall standard deviation (SD) of the portfolio.

3. Sharpe ratio (SR): a common measure of risk-adjusted return. Unfortunately, this penalises both upside and downside volatility. Some strategies with strong annualised returns may have lower Sharpe ratios due to erratic, yet positive return profiles.

4. Maximum drawdown (MD): an alternative penalty function capturing the maximum peak to trough trading loss.

5. Sortino ratio (SoR): an alternative measure of risk-adjusted return that only penalises downside deviation in the denominator.

6. Calmar ratio (CR): a measure of risk-adjusted returns that penalises the maximum realised drawdown over some candidate investment period.

4.1. Model performance: 2008-2020

Implementing our trading strategy between January 2008 and April 2020 would have been highly successful for both risk-averse and risk-on investors. Seen in Table 6 and Figure 3, the strategy consistently outperformed the S&P 500 index, and overall generated annualised returns of 11%. The S&P 500 returned 5.4% while the static baseline strategy returned 6.3%. The strategy clearly generates alpha by its dynamic nature, automatically detecting market regimes and allocating capital successfully. This entire period can broadly be characterised as a bull market, and yet features several severe market shocks; the strategy's consistent performance demonstrates its robustness to varied market dynamics. Figure 4 shows the positions held by the strategy.

Strategy	AR	SD	SR	MD	SoR	CR
Hold SPY	0.054	0.21	0.36	0.53	0.50	0.10
Hold GLD	0.054	0.18	0.48	0.46	0.54	0.12
Baseline	0.063	0.14	0.52	0.33	0.73	0.19
Dynamic	**0.11**	**0.16**	**0.72**	**0.33**	**1.03**	**0.33**

Table 6: Validation metrics: January 2008 - April 2020

Figure 3: Cumulative returns: January 2008 - April 2020

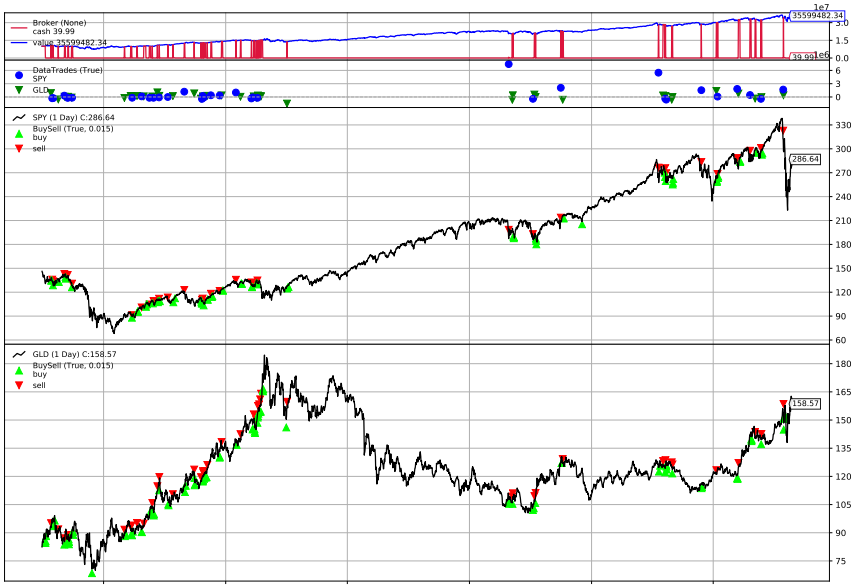


Figure 4: Positions held by dynamic strategy: January 2008 - April 2020

Of the four strategies compared, our dynamic trading strategy has the best annualized returns, Sharpe ratio, Sortino ratio and Calmar ratio, and lowest drawdown. It has the second lowest standard deviation of 0.16, close to the baseline static strategy's 0.14. The most significant component of the Sharpe ratio's performance comes from strong annualised returns; the increased upside volatility is the main contributor to the standard deviation. Indeed, our strategy's Sortino ratio is 3 times greater than that of the S&P 500; this confirms that a significant degree of the penalty in the standard deviation and Sharpe ratio is generated from upside returns. That is, the strong annualised returns of our trading strategy are generated in a relatively volatile manner. This is unsurprising, given that the strategy generates performance due to market timing.

4.2. Detailed analysis of performance over time

In this section, we describe the performance in detail over various time periods, particularly during market crises. Note: while we have reported our findings over one period 2008-2020, in fact four separate learning and evaluation procedures have been performed. All four periods were successful for our strategy, visible in Figure 3.

First, the strategy performs well during the GFC. Our strategy generates the second best returns during the GFC, surpassed only by gold. During the GFC, gold provided extraordinary returns for investors who invested prior to or during the crisis. After incurring a sharp drawdown, our strategy reallocates capital from S&P 500 into gold and consequently outperforms equity markets until late 2011.

Next, the market experienced significant drawdown in December 2018. Given the brevity of this drawdown, our trading strategy is unable to reallocate capital away from the S&P 500 into gold fast enough to meaningfully reduce the strategy's drawdown. After all, our strategy is predicated on identifying regimes, and allocating capital when new data are identified as similar to past phenomena. It reflects the delicate balance in the look back length n. If it were too long, trading decisions would be made too slowly; if it were too short, trading decisions would be made too frivolously.

The final significant market crisis during our window of analysis is the market turbulence associated with COVID-19. Our strategy performs extremely well during this period. Although the strategy does experience losses in late January and February 2020, capital is reallocated toward gold and strategy returns recover quickly. In fact, the cumulative returns of the strategy are no lower than the previous high, prior to the COVID-19 crisis. In comparison, the S&P 500 and our static baseline strategy suffered more significant drawdowns, and have failed to return to prior high watermark levels.

During the four 4-year windows that make up the 2008-2020 experiment, the optimal look back length n changes as follows. For the four windows, the optimal chosen n is 13, 13, 18 and 16 for 2004-2008, 2008-2012, 2012-2016 and 2016-2020 respectively. This suggests that continually updating the look back length is important, due to the dynamic nature of markets. Note that the longest

look back length is during 2012-2016, a bull market period with the greatest consistency and least volatility in the return profile. This suggests that regimes were more persistent and possibly easier to identify during the 2012-2016 period.

5. Conclusion

We have shown that our new methodology for clustering volatility regimes is a useful tool for making inferences on financial time series and for designing trading strategies. Results on both synthetic and real data are promising, with good validation scores and significant simplification of the time series. These findings help support the work by [1, 3] who contributed to the idea of discrete changes in volatility regimes. Moreover, while these models generally select the number of regimes to begin with, we have have determined the number of clusters, and showed this is overwhelmingly 2 in practice. And yet, our method is flexible enough to detect greater numbers of regimes if clearly present, such as piecewise autoregressive models. Our method fits well with others in the literature, as our determined number of volatility regimes can then be used in an alternative regime switching model, which generally requires this number to be set *a priori*.

Our dynamic trading strategy performs well at avoiding periods of significant volatility and drawdown, and performs substantially better than the S&P 500, in various market conditions. Our method continually updates its distributions and parameters, reflecting the need for ongoing learning of market conditions and volatility structure. Our method is also flexible, with several natural alternatives one could adopt. For instance, one could switch from SPY to cash as an alternative safe haven asset, replacing gold. Our methodology could also be combined with other statistical or machine learning methods in the literature. For example, instead of a static safe haven class to which the strategy flees in times of volatility, one could use a learned allocation of low beta assets as an evolving safe haven.

Acknowledgements

Many thanks to Alex Judge for helpful comments and suggestions.

Appendix A. Details of change point detection algorithm

Appendix A.1. General change point detection framework

First, we outline the change point detection framework in greatest generality. A sequence of observations $x_1, x_2, ..., x_n$ are drawn from random variables $X_1, X_2, ..., X_n$. We wish to determine points $\tau_1, ..., \tau_m$ at which the distributions change. One always assumes that the underlying random variables are independent and identically distributed between change points. One can summarize this with the following notation, following Ross [14]:

$$X_i \sim \begin{cases} F_0 \text{ if } i \leq \tau_1 \\ F_1 \text{ if } \tau_1 < i \leq \tau_2 \\ F_2 \text{ if } \tau_2 < i \leq \tau_3, \\ \dots \end{cases}$$

That is, one assumes X_i is a random sampling of a different distribution over each time period $[\tau_i, \tau_{i+1}]$. In order to meet the apparently restrictive assumption of independence of the data, one must usually perform an appropriate transformation of the data. The log quotient transformation, which yields the log returns from the closing price data, is one such transformation [24].

Appendix A.2. Rank of observations and Mood Test

Ross [25] points out the fact that log returns often exhibit heavy tailed behaviour. As a result, a non-parametric test is needed to detect change points that do not *a priori* assume the distribution of the data. The *rank test* is one such test. Suppose there are two samples of observations from unknown distributions $A = \{r_{1,1} =, r_{1,2}, ..., r_{1,m}\}$ and $B = \{r_{2,1} =, r_{2,2}, ..., r_{2,n}\}$. Define the *rank* of an observation $r \in A \cup B$ as follows:

$$\text{rank}(r) = \sum_j^m \mathbb{1}_{(r \geq r_{1,j})} + \sum_j^n \mathbb{1}_{(r \geq r_{2,j})} = \#\{s \in A \cup B : r \geq s\}$$

A larger rank indicates a higher positioning in the ordering of the elements of A and B. If both sets of samples have the same distribution, the median rank among $\{\text{rank}(r) : r \in A \cup B\}$ is $\frac{1}{2}(n + m + 1)$. In this case, one would assume that both sets have a near equal split of the ranks.

The *Mood test* determines the extent that each observation's rank differs from the median rank, thereby detecting differences in the distributions' variance. If the samples have different variances, then one set of samples would have more extreme values than the other, which means the ranks would not be even between the two sets. Specifically, the test statistic is as follows:

$$M'_{m,n} = \sum_{i=1}^{m} (\text{rank}(r_{1,i}) - (m + n + 1)/2)^2$$

This is appropriately normalized:

$$\begin{aligned} N &= m + n \\ \mu_{M'} &= \frac{1}{12} m(N^2 - 1) \\ \sigma_{M'} &= \frac{1}{180} mn(N+1)(N^2 - 4) \\ M_{mn} &= \frac{1}{\sigma_{M'}}(|M' - \mu_{M'}|) \end{aligned}$$

If M_{mn} is greater than some threshold h, we reject the null hypothesis that the distributions have the same variance, and conclude they have different variances. As depicted in Appendix B, the log return time series are tail heavy but strongly mean and median centred. Thus, the Mood test reliably detects changes in the variance without being affected by changes in the median. Compare Sections 4 and 5 of [11] for this distinction.

Appendix A.3. CPM algorithm

Ross' CPM algorithm [14] works by feeding in one data point at a time. When a change point τ is detected, the algorithm restarts and proceeds from that point, so it suffices to describe how the algorithm determines its very first change point.

Suppose $x_1, ..., x_N$ is a sequence for which no change point has been detected. For each $m = 1, 2, ..., N$ define $n = N - m$, mirroring the notation of Appendix A.2, and compute the Mood test statistic $M_{m,n}$. If the maximum among these, $M_N = \max_{m+n=N} M_{m,n}$, exceeds a threshold parameter h_N, we declare a change point in the variance has occurred at $\hat{\tau} = \text{argmax}_m M_{m,n}$. If the maximum such test statistic does not exceed the threshold parameter, feed in the next data point x_{N+1} and continue. Note if a change point $\hat{\tau} = m$ is detected at time N, there has been a delay of n units in its detection. This delay is necessary for the algorithm to examine data points on each side of the change point. The algorithm then restarts from the change point $\hat{\tau}$.

In our implementation of the algorithm, we always read in at least 30 values before looking for another change point, so that all stationary periods have length at least 30. We choose our parameters h in order to manage the number of false positives (Type I errors). Given an acceptability threshold α, the following equations specify that this error should remain constant over time:

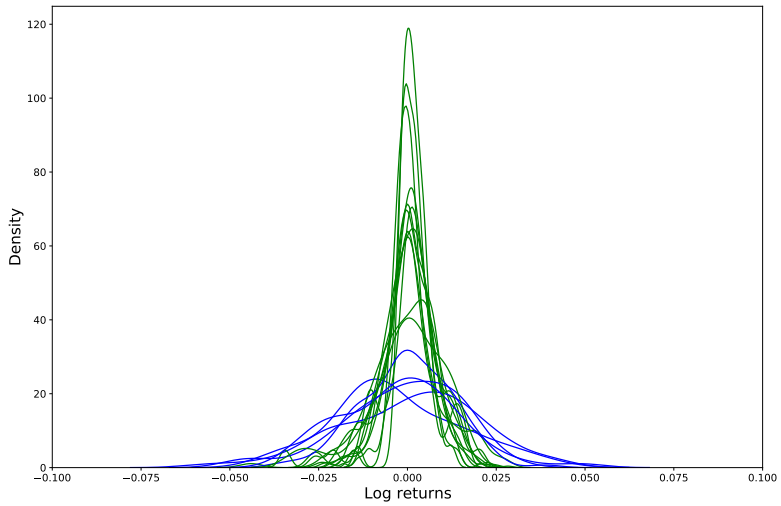
$$\begin{aligned} P(M_1 > h_1) &= \alpha \\ P(M_t > h_t | M_{t-1} \leq h_{t-1}, ..., M_1 \leq h_1) &= \alpha \end{aligned}$$

In the event that no change point exists, a false positive will nonetheless be detected at time $1/\alpha$ on average. This quantity is the average run length parameter ARL_0 that is passed to CPM, which in term calculates the appropriate choice of h_t. In this case ARL_0 is set to 10,000.

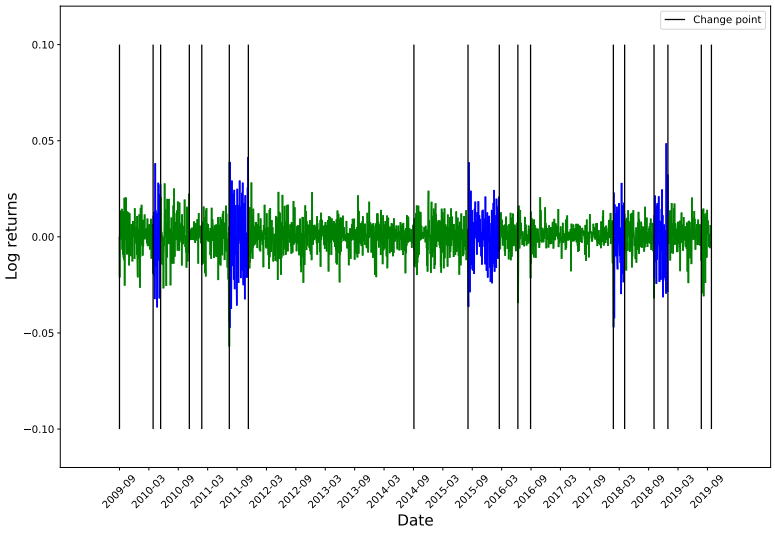
Appendix B. Plots

What follows are plots pertaining to the body of the paper. First, the complete set of results for Section 3 can be found below. Figures B.5, B.6, B.7, B.8 contain depict the clustered distributions and volatility regimes of the Dow Jones, Nikkei, FTSE and MSCI indices respectively. Figures B.9, B.10,B.11, B.12, B.13 do so for individual firms Microsoft, Apple, Amazon, Alphabet and Berkshire Hathway Class A respectively. Figures B.14, B.15, B.16, B.17, B.18 depict popular ETFs RYT, GLD, XLF, IJS and DRGW respectively. Figures B.19, B.20, B.21, B.22, B.23 depict the distributions and regimes for the JPY/USD, AUD/USD, EUR/USD, GBP/USD and NZD/AUD.

Note all distribution plots are strongly centred in mean and median about zero. This is an important technical point for the Mood test to work correctly to detect changes in variance.

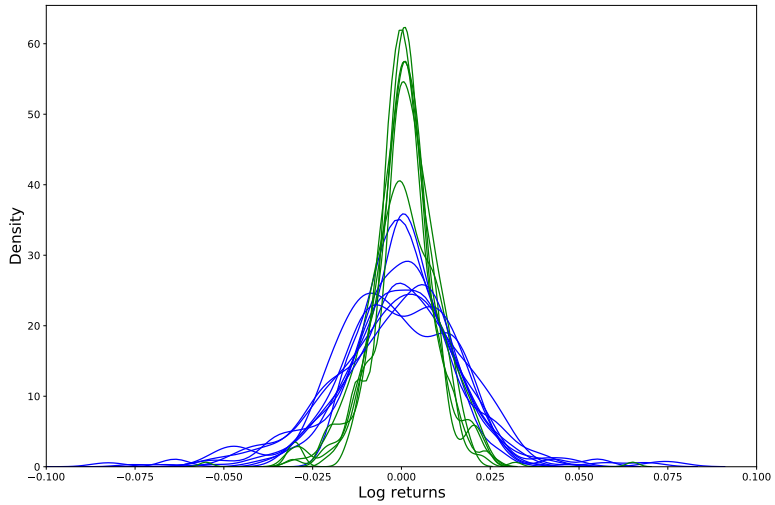


(a) Dow Jones clustered distributions

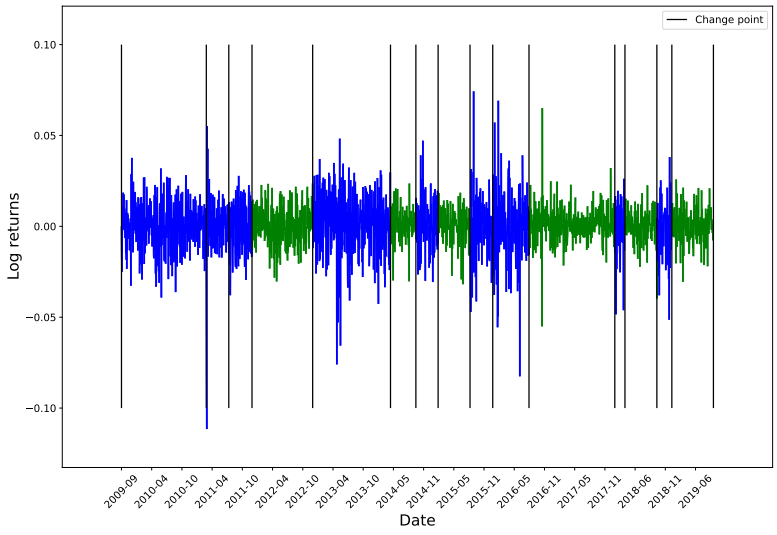


(b) Dow Jones volatility regimes

Figure B.5: Dow Jones results

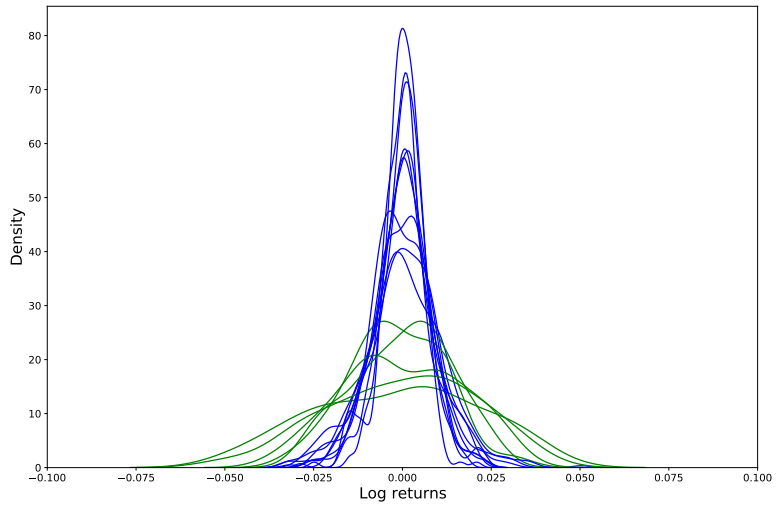


(a) Nikkei clustered distributions

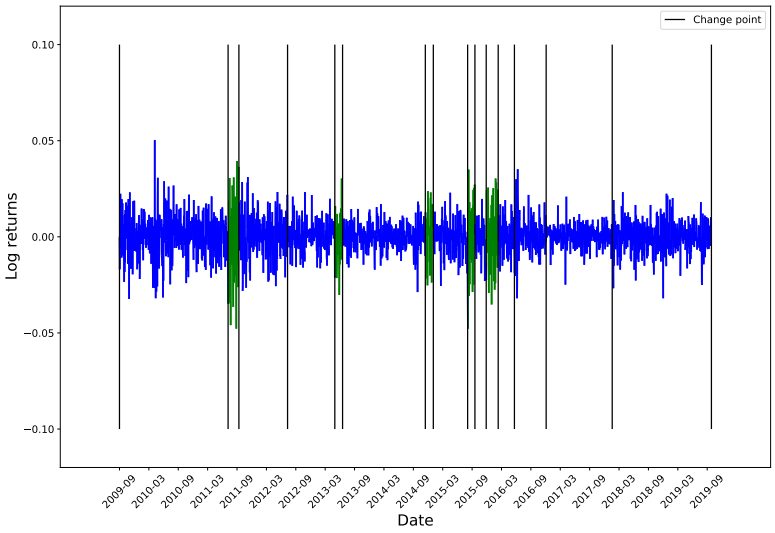


(b) Nikkei volatility regimes

Figure B.6: Nikkei 225

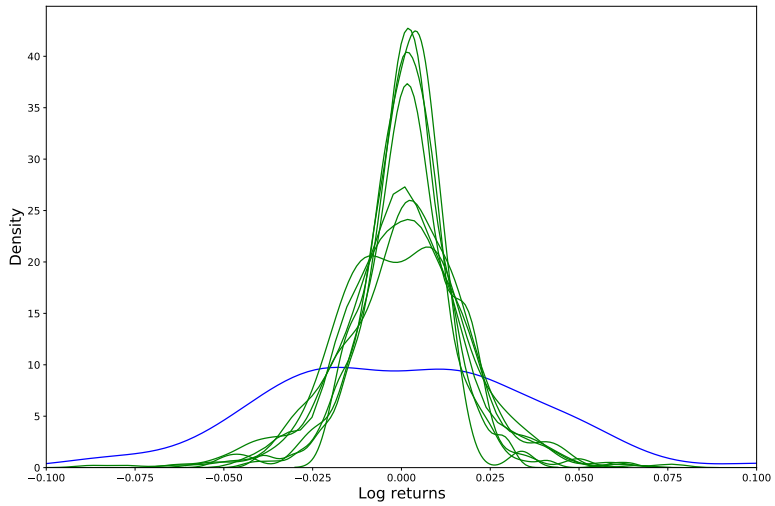


(a) FTSE clustered distributions

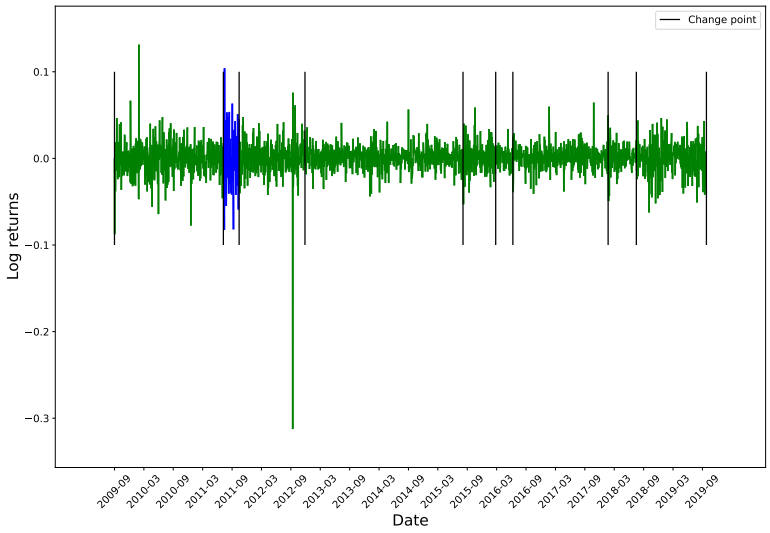


(b) FTSE volatility regimes

Figure B.7: FTSE 100

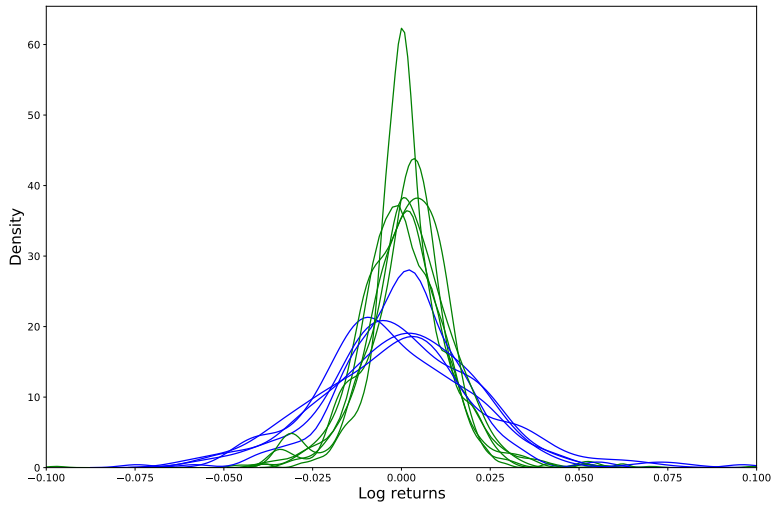


(a) MSCI clustered distributions

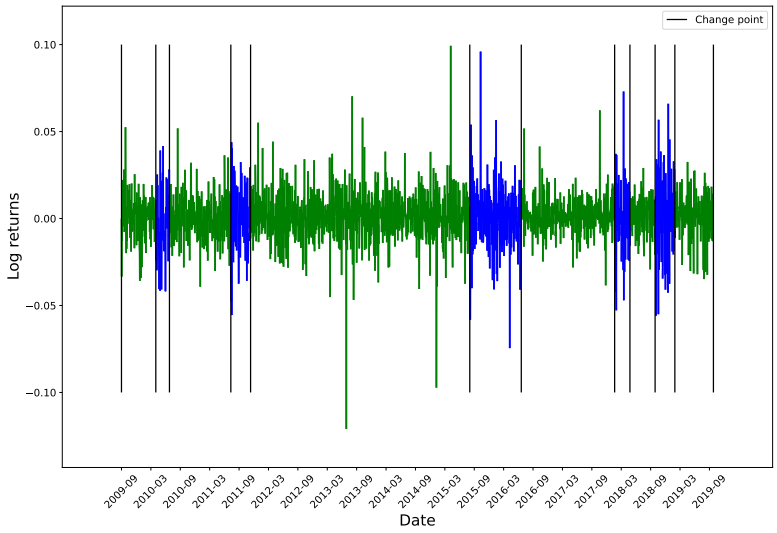


(b) MSCI volatility regimes

Figure B.8: MSCI world index

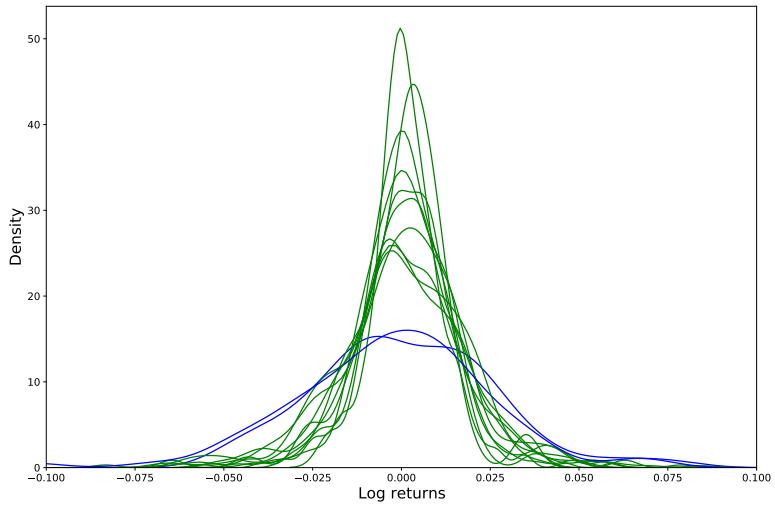


(a) MSFT clustered distributions

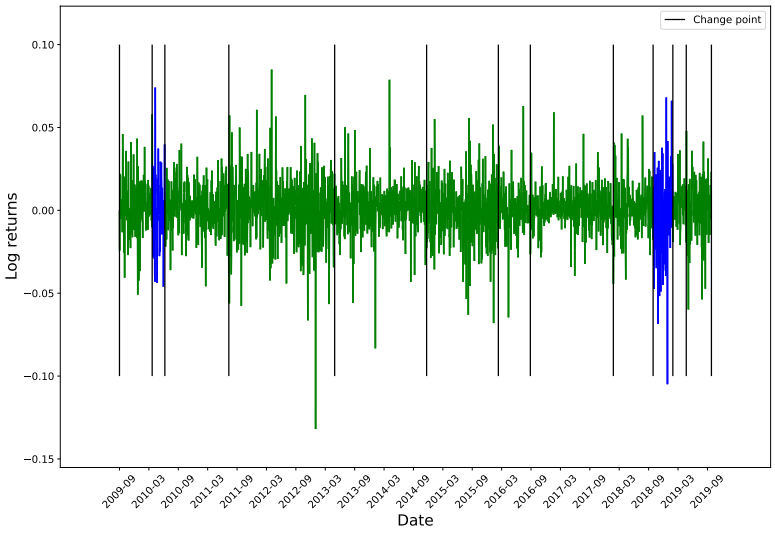


(b) MSFT volatility regimes

Figure B.9: Microsoft

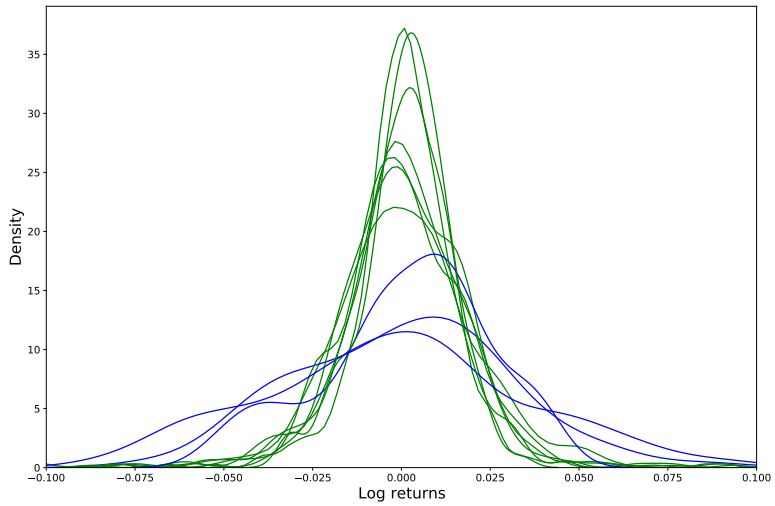


(a) AAPL clustered distributions

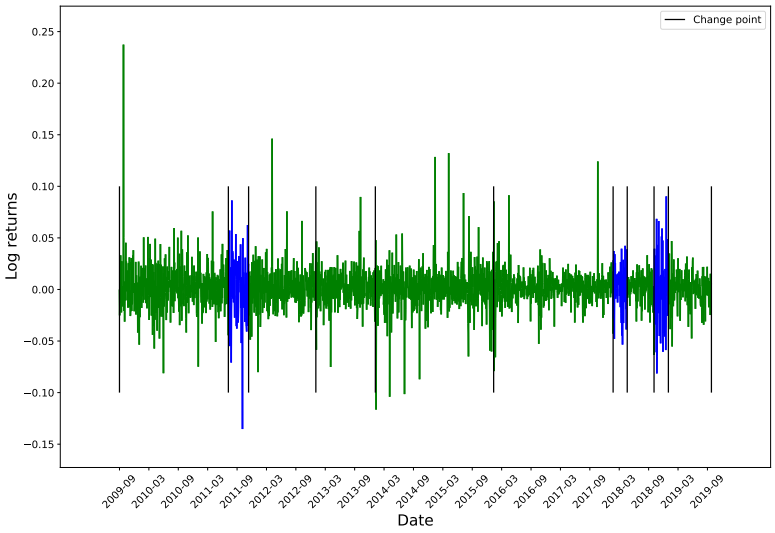


(b) AAPL volatility regimes

Figure B.10: Apple

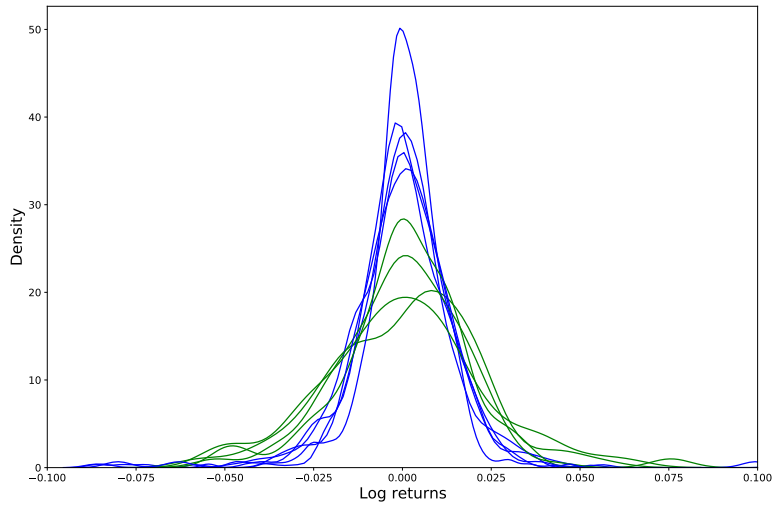


(a) AMZN clustered distributions

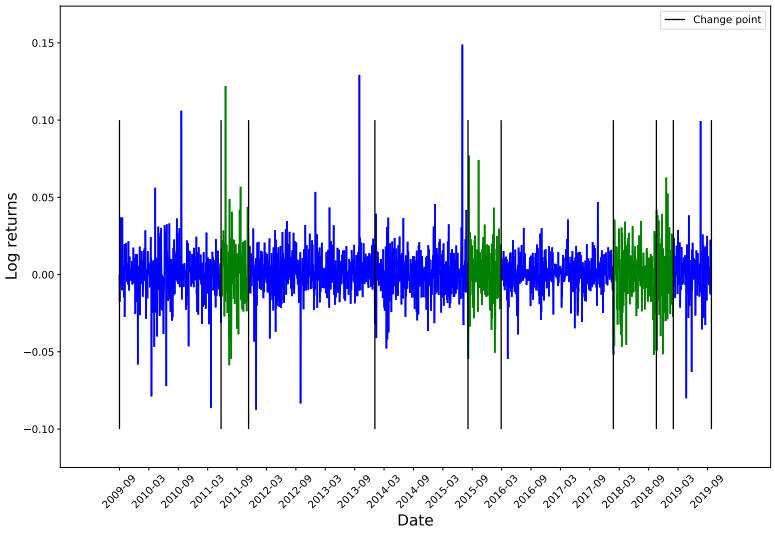


(b) AMZN volatility regimes

Figure B.11: Amazon

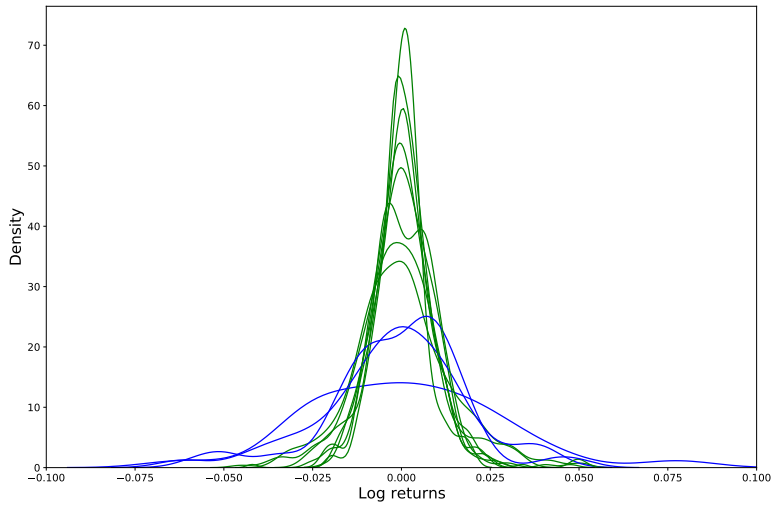


(a) GOOG clustered distributions

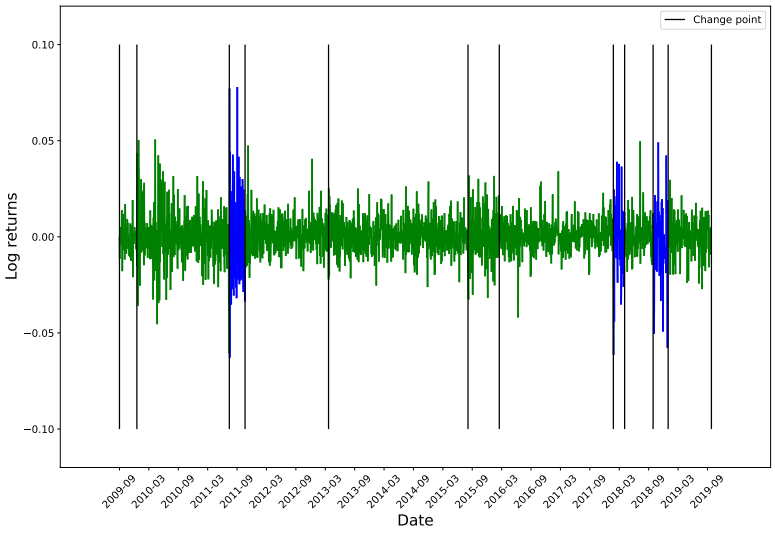


(b) GOOG volatility regimes

Figure B.12: Alphabet

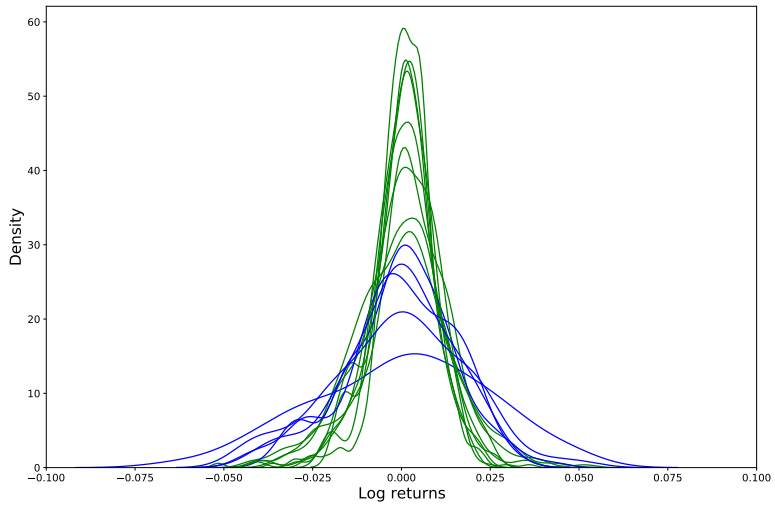


(a) BRK-A clustered distributions

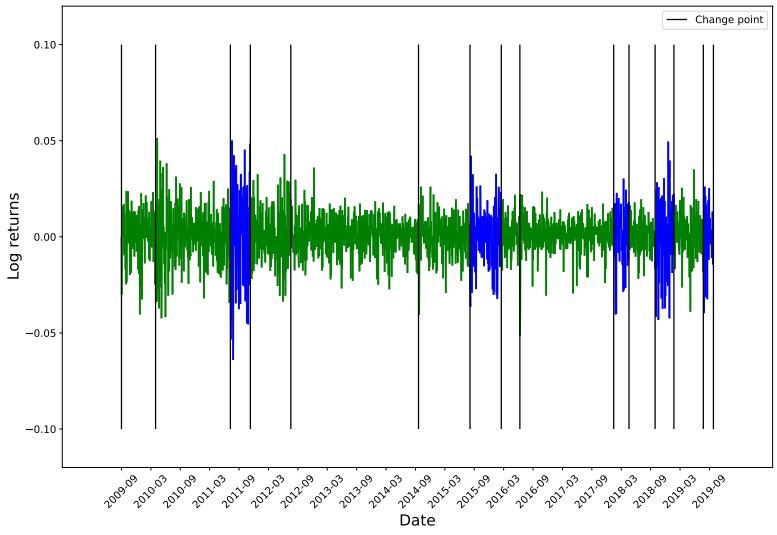


(b) BRK-A volatility regimes

Figure B.13: Berkshire Hathaway Class A

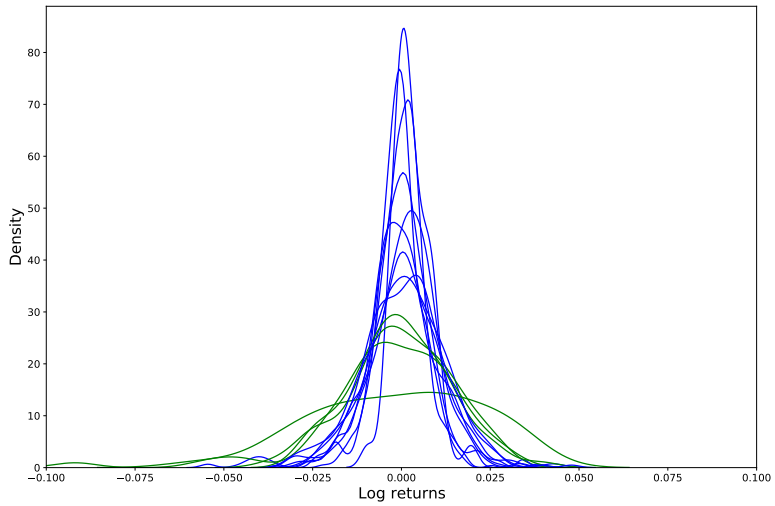


(a) RYT clustered distributions

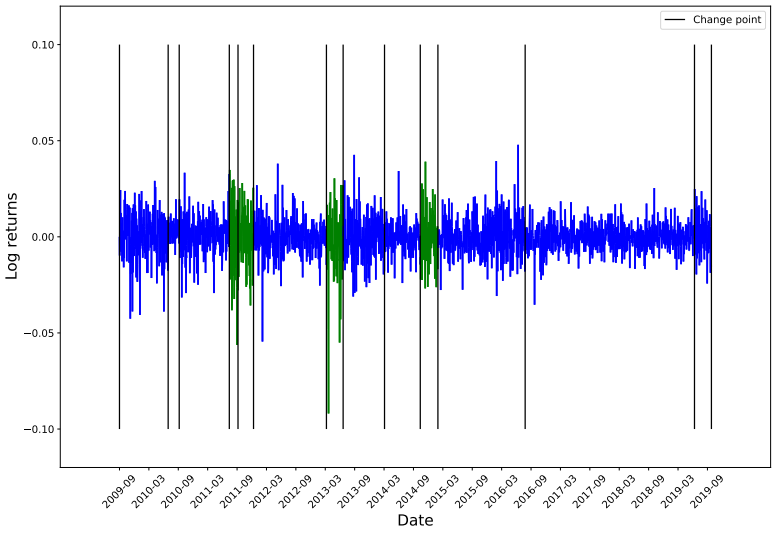


(b) RYT volatility regimes

Figure B.14: RYT: Invesco S&P 500 Equal Weight Technology ETF

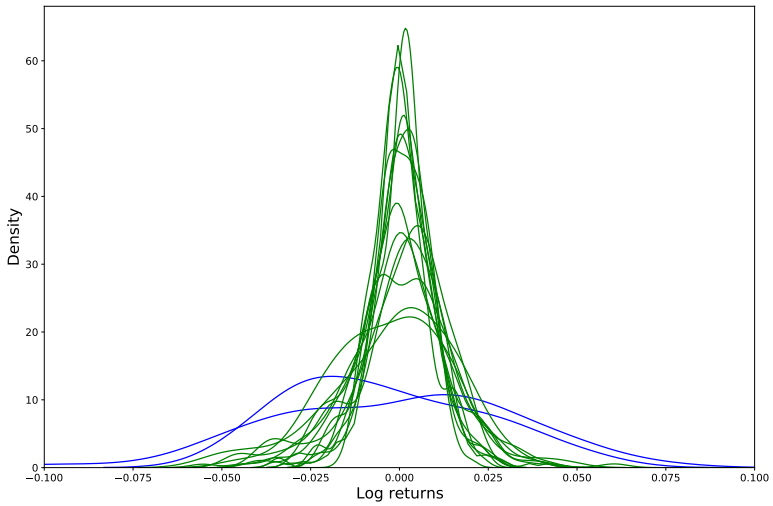


(a) GLD clustered distributions

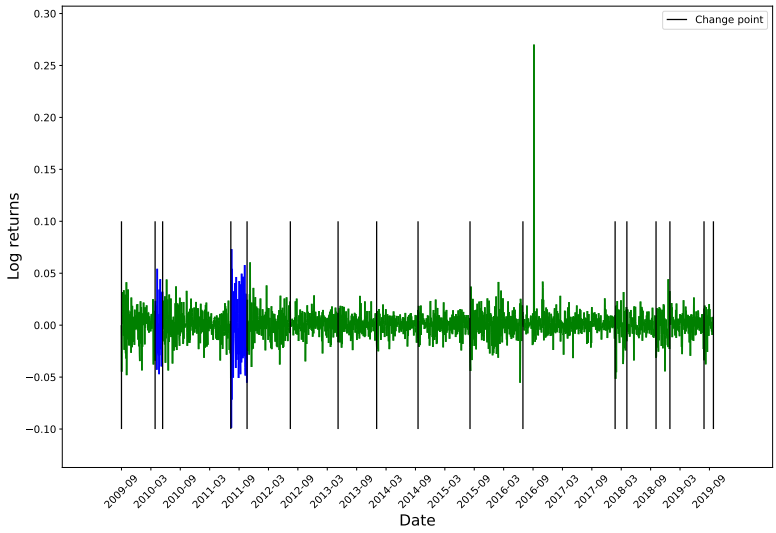


(b) GLD volatility regimes

Figure B.15: GLD: SPDR Gold Shares

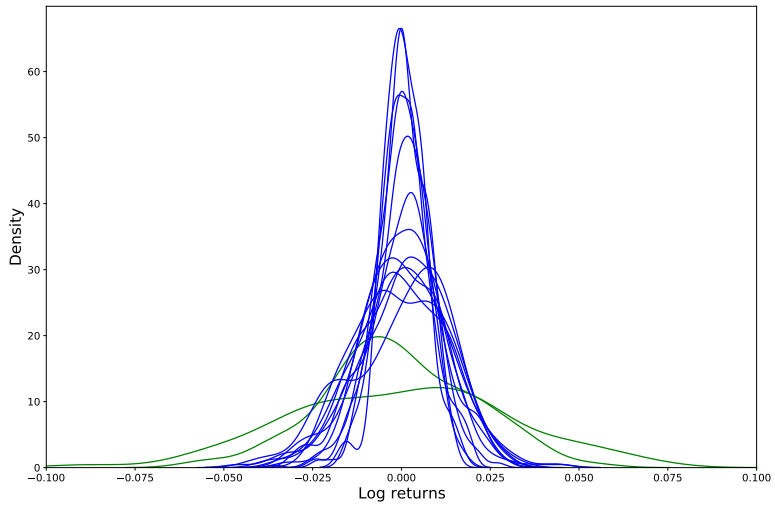


(a) XLF clustered distributions

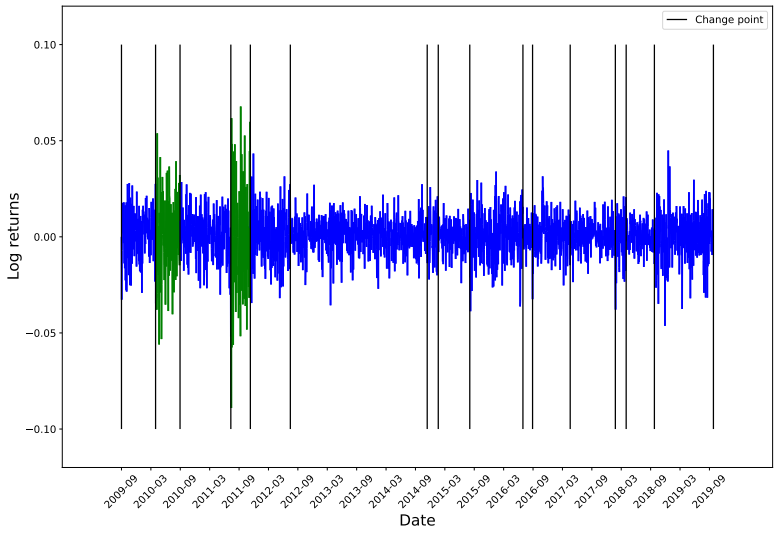


(b) XLF volatility regimes

Figure B.16: XLF: Financial Select Sector SPDR Fund

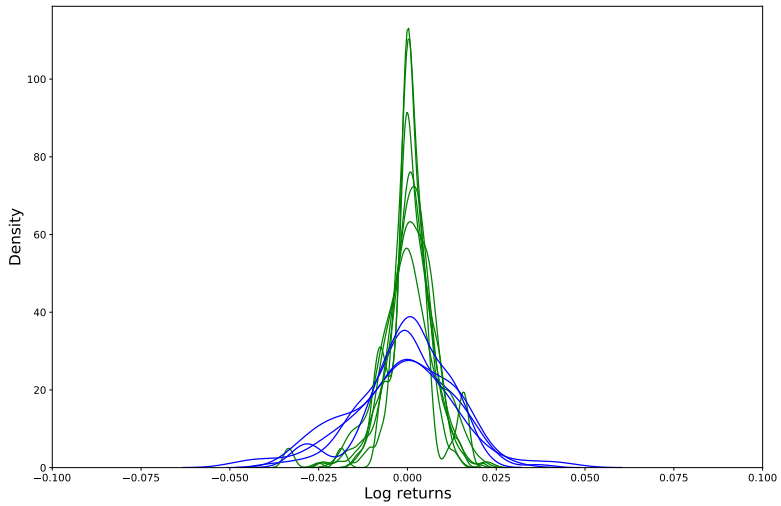


(a) IJS clustered distributions

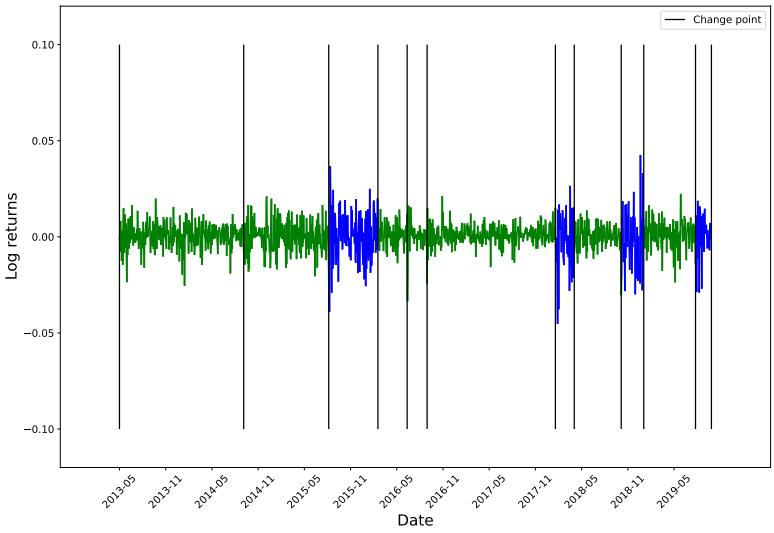


(b) IJS volatility regimes

Figure B.17: IJS: iShares SP Small-Cap 600

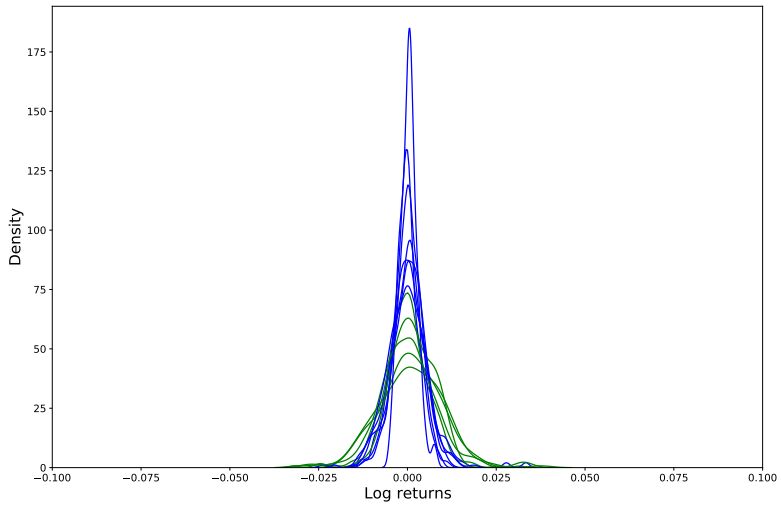


(a) DGRW clustered distributions

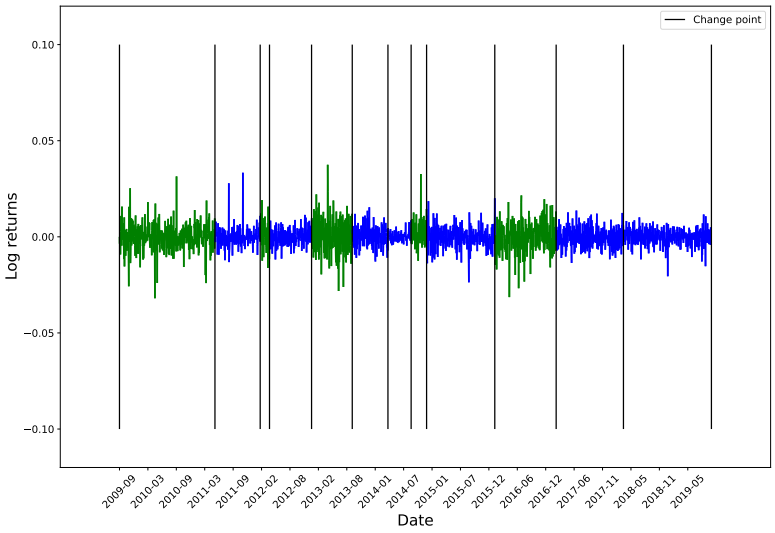


(b) DRGW volatility regimes

Figure B.18: DRGW: WisdomTree U.S. Quality Dividend Growth Fund

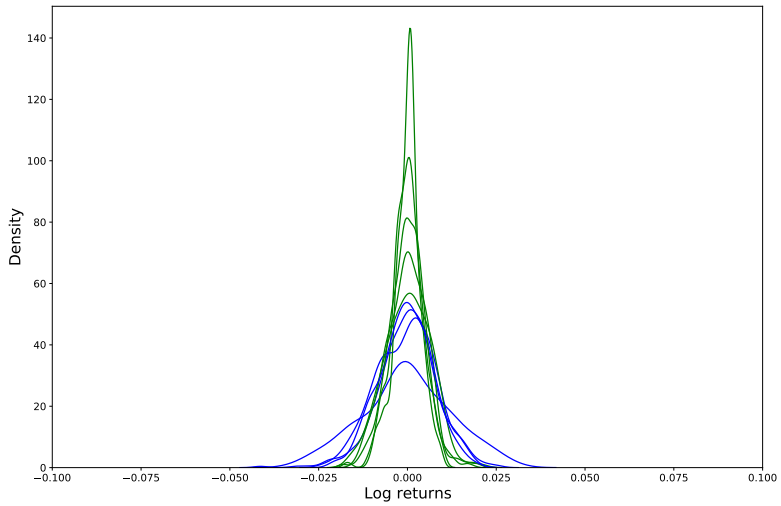


(a) JPY clustered distributions

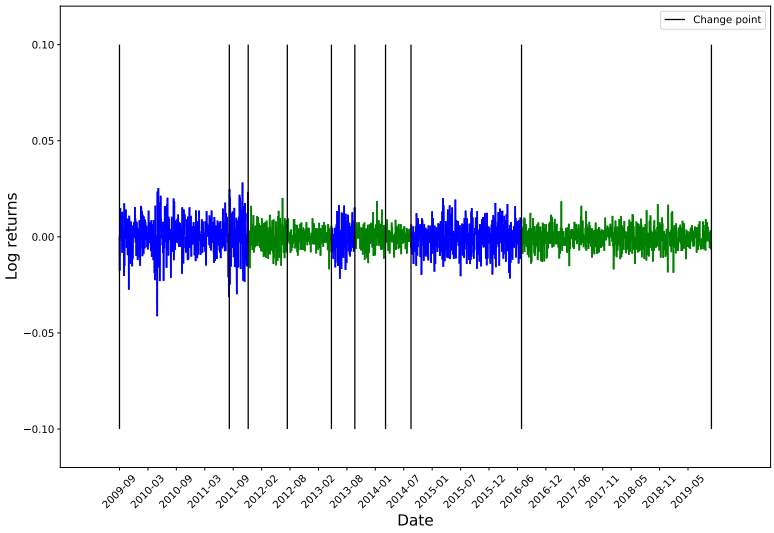


(b) JPY volatility regimes

Figure B.19: USD/JPY

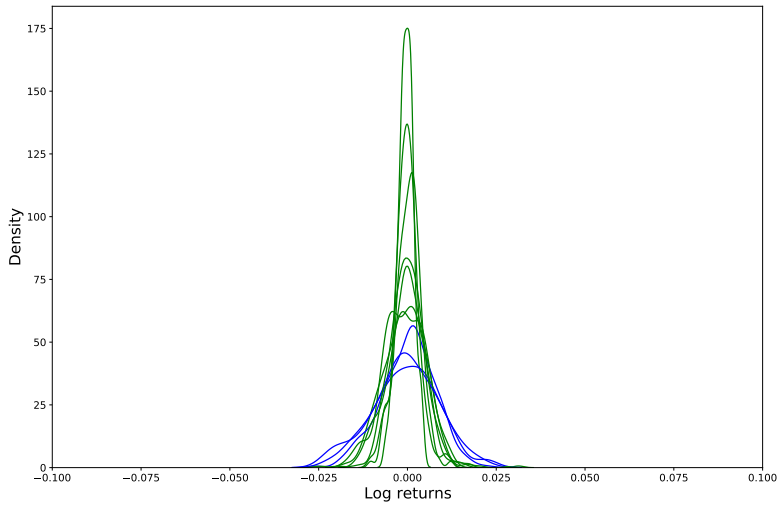


(a) AUD clustered distributions

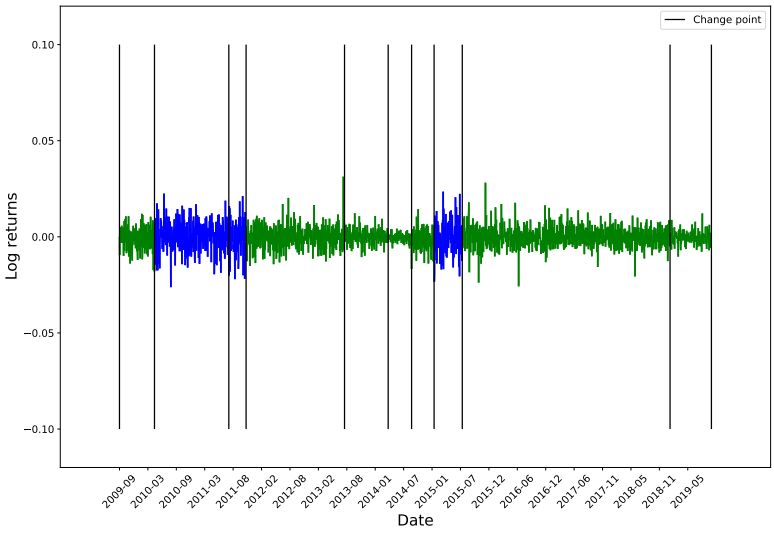


(b) AUD volatility regimes

Figure B.20: AUD/USD

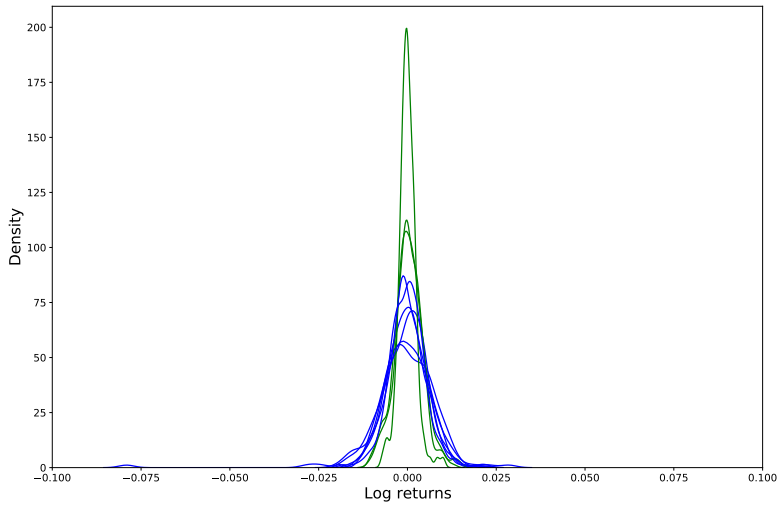


(a) EUR clustered distributions

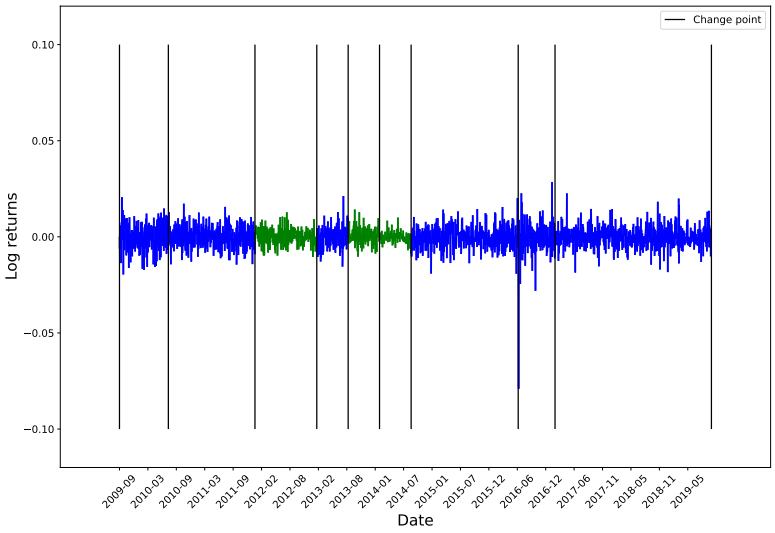


(b) EUR volatility regimes

Figure B.21: EUR/USD

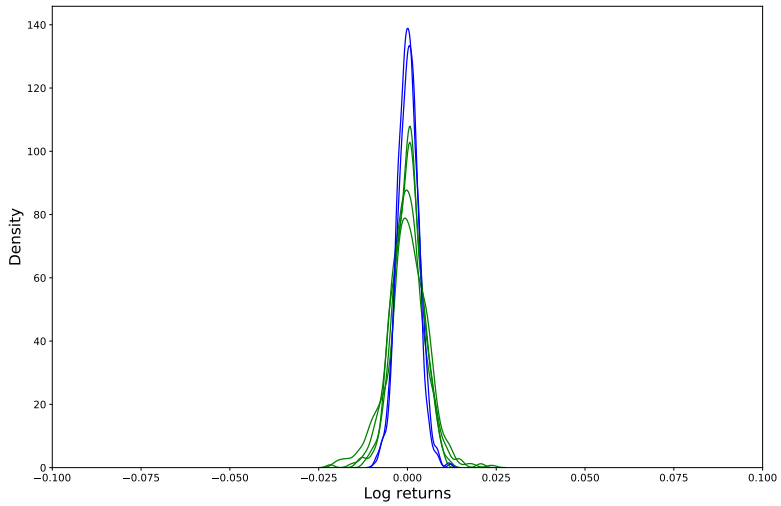


(a) GBP clustered distributions

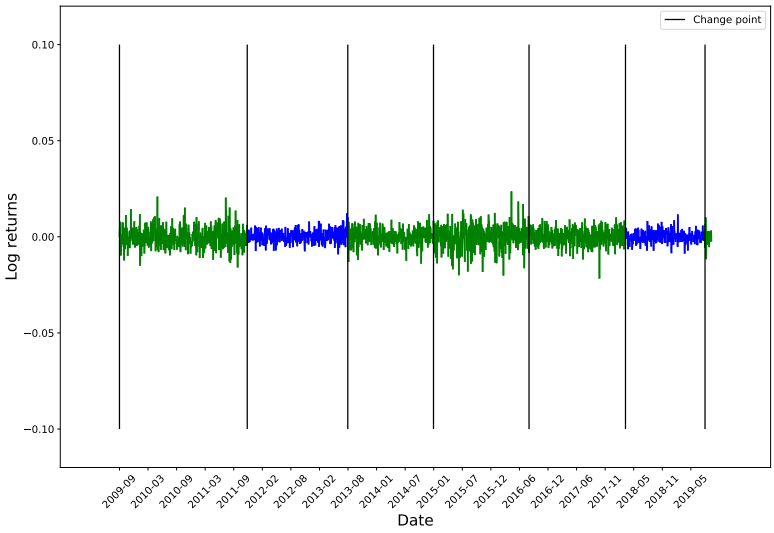


(b) GBP volatility regimes

Figure B.22: GBP/USD



(a) NZD clustered distributions



(b) NZD volatility regimes

Figure B.23: NZD/AUD

References

[1] J. D. Hamilton, A new approach to the economic analysis of nonstationary time series and the business cycle, Econometrica 57 (1989) 357–384. doi:10.2307/1912559.

[2] M. Lavielle, G. Teyssière, Adaptive detection of multiple change-points in asset price volatility, in: Long Memory in Economics, Springer Berlin Heidelberg, 2007, pp. 129–156. doi:10.1007/978-3-540-34625-8_5.

[3] C. G. Lamoureux, W. D. Lastrapes, Persistence in variance, structural change, and the GARCH model, Journal of Business & Economic Statistics 8 (1990) 225–234. doi:10.2307/1391985.

[4] D. Shah, H. Isah, F. Zulkernine, Stock market analysis: A review and taxonomy of prediction techniques, International Journal of Financial Studies 7 (2019) 26. doi:10.3390/ijfs7020026.

[5] R. F. Engle, Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation, Econometrica 50 (1982) 987–1007. doi:10.2307/1912773.

[6] T. Bollerslev, Generalized autoregressive conditional heteroskedasticity, Journal of Econometrics 31 (1986) 307–327. doi:10.1016/0304-4076(86)90063-1.

[7] D. M. Guillaume, M. M. Dacorogna, R. R. Davé, U. A. Müller, R. B. Olsen, O. V. Pictet, From the bird's eye to the microscope: A survey of new stylized facts of the intra-daily foreign exchange markets, Finance and Stochastics 1 (1997) 95–129. doi:10.1007/s007800050018.

[8] F. Klaassen, Improving GARCH volatility forecasts with regime-switching GARCH, Empirical Economics 27 (2002) 363–394. doi:10.1007/s001810100100.

[9] D. M. Hawkins, Testing a sequence of observations for a shift in location, Journal of the American Statistical Association 72 (1977) 180–186. doi:10.1080/01621459.1977.10479935.

[10] D. M. Hawkins, K. D. Zamba, A change-point model for a shift in variance, Journal of Quality Technology 37 (2005) 21–31. doi:10.1080/00224065.2005.11980297.

[11] A. M. Mood, On the asymptotic efficiency of certain nonparametric two-sample tests, The Annals of Mathematical Statistics 25 (1954) 514–522. doi:10.1214/aoms/1177728719.

[12] P. Nystrup, B. W. Hansen, H. Madsen, E. Lindström, Detecting change points in VIX and S&P 500: A new approach to dynamic asset allocation, Journal of Asset Management 17 (2016) 361–374. doi:10.1057/jam.2016.12.

[13] R. Dahlhaus, Fitting time series models to nonstationary processes, The Annals of Statistics 25 (1997) 1–37. doi:10.1214/aos/1034276620.

[14] G. J. Ross, Parametric and nonparametric sequential change detection in R: The cpm package, Journal of Statistical Software, Articles 66 (2015) 1–20. URL: https://www.jstatsoft.org/v066/i03. doi:10.18637/jss.v066.i03.

[15] T. G. Andersen, T. Bollerslev, F. X. Diebold, H. Ebens, The distribution of realized stock return volatility, Journal of Financial Economics 61 (2001) 43–76. doi:10.1016/s0304-405x(01)00055-1.

[16] B. W. Silverman, Density Estimation for Statistics and Data Analysis, Chapman and Hall, 1986. doi:10.1201/9781315140919.

[17] E. M. Stein, R. Shakarchi, Real Analysis: Measure Theory, Integration, and Hilbert Spaces, Princeton University Press, 2005. doi:10.1017/s0025557200181343.

[18] E. del Barrio, E. Giné, C. Matrán, Central limit theorems for the wasserstein distance between the empirical and the true distributions, The Annals of Probability 27 (1999) 1009–1071. doi:10.1214/aop/1022677394.

[19] U. von Luxburg, A tutorial on spectral clustering, Statistics and Computing 17 (2007) 395–416. doi:10.1007/s11222-007-9033-z.

[20] P. J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, Journal of Computational and Applied Mathematics 20 (1987) 53–65. doi:10.1016/0377-0427(87)90125-7.

[21] L. Kaufman, P. J. Rousseeuw, Finding Groups in Data, John Wiley & Sons, Inc., 1990. doi:10.1002/9780470316801.

[22] A. Korotayev, S. Tsirel, A spectral analysis of world GDP dynamics: Kondratieff waves, Kuznets swings, Juglar and Kitchin cycles in global economic development, and the 2008–2009 economic crisis, Structure and Dynamics : e-Journal of Anthropological and Related Sciences 4 (2010).

[23] M. Gärtner, K. W. Wellershoff, Is there an election cycle in American stock returns?, International Review of Economics & Finance 4 (1995) 387–410. doi:10.1016/1059-0560(95)90036-5.

[24] F. Gustafsson, Adaptive Filtering and Change Detection, John Wiley & Sons, Ltd, 2001. doi:10.1002/0470841613.

[25] G. J. Ross, Modelling financial volatility in the presence of abrupt changes, Physica A: Statistical Mechanics and its Applications 392 (2013) 350–360. doi:10.1016/j.physa.2012.08.015.