

CLUSTERING MARKET REGIMES USING THE WASSERSTEIN DISTANCE

B. HORVATH¹, Z. ISSA^{*2}, AND A. MUGURUZA³

ABSTRACT. The problem of rapid and automated detection of distinct market regimes is a topic of great interest to financial mathematicians and practitioners alike. In this paper, we outline an unsupervised learning algorithm for clustering financial time-series into a suitable number of temporal segments (market regimes). As a special case of the above, we develop a robust algorithm that automates the process of classifying market regimes. The method is robust in the sense that it does not depend on modelling assumptions of the underlying time series as our experiments with real datasets show. This method – dubbed the Wasserstein k -means algorithm – frames such a problem as one on the space of probability measures with finite p^{th} moment, in terms of the p -Wasserstein distance between (empirical) distributions. We compare our WK-means approach with a more traditional clustering algorithms by studying the so-called maximum mean discrepancy scores between, and within clusters. In both cases it is shown that the WK-means algorithm vastly outperforms all considered competitor approaches. We demonstrate the performance of all approaches both in a controlled environment on synthetic data, and on real data.

Keywords: Regime clustering, unsupervised learning, optimal transport

Short title: Clustering market regimes using the Wasserstein distance

Word count: 7954 (excluding appendices)

Number of figures and tables: 35 figures, 4 tables

CONTENTS

1. Introduction	2
2. k -means on the space of distributions	9
3. Methodology and numerical results	11
4. Conclusion	27
References	29
Appendix	32

^{*}Corresponding author.

¹Department of Mathematics, Oxford University and the Oxford Man Institute, Oxford, United Kingdom, OX2 6GG, blanka.horvath@maths.ox.ac.uk

²Department of Mathematics, King's College London, Strand, London, United Kingdom, WC2R 2LS, zacharia.issa@kcl.ac.uk

³Kaiju Capital Management, 3rd Floor, Palm Grove House, Road Town, Tortola, British Virgin Islands VG1110, aitor.muguruza-gonzalez@kaiju.ai

KEY MESSAGES

The main findings in this paper are as follows:

- (1) We develop and apply a modification of the k -means clustering algorithm to financial data using the Wasserstein distance and Wasserstein barycenter
- (2) We qualitatively compare our model to two classical approaches and show that it outperforms them, especially when returns are non-Gaussian
- (3) We quantitatively validate our approach using a wide variety of methods, including some classical, and a maximum mean discrepancy approach, validating qualitative results

1. INTRODUCTION

Time series data derived from asset returns are known to exhibit certain properties, termed *stylised facts*, that are ubiquitous across asset classes. For example, it is well-understood that return series are non-stationary in the strong sense, and exhibit volatility clustering (for a full recount, see Cont (2001)). In particular, understanding the heteroskedastic nature of financial time series data is relevant for market practitioners. An observed sequence of asset returns (or, for multiple assets, a tuple of sequences) tend to exhibit periods of similar behaviour, followed by those that indicate a significantly different underlying distribution. Such periods are often referred to as *market regimes*. Our interest in swiftly and accurately detecting changes in market regimes is motivated by a multitude of financial applications (both classical and modern): Most naturally, an accurate detection of shifts in market behaviour is tantamount for making optimised investment decisions or trading strategies. But also within the arena of recent, deep learning-based methods for pricing hedging and market generation, the detection of significant shifts in market behaviour is a central tool for their model governance, since it serves as an indicator for the need to retrain the ML-model, see Horvath, Muguruza, and Tomas (2019); Buehler, Horvath, Lyons, Perez Arribas, and Wood (2020). Henceforth, we call the task of finding an effective way of grouping different regimes the *market regime clustering problem (MRCP)*.

In this paper, we propose a methodology to classify segments of the historical evolution of market returns into distinct regimes. We do so by devising a modified, versatile version of the classical k -means clustering algorithm to group distributions of asset returns into regimes, which exhibit a higher degree of homogeneity. The way modify the classical algorithm is twofold: Firstly, by a shift of perspective, we consider the clustering problem as one on the space of distributions with finite p^{th} moment, as opposed to one on Euclidean space. Secondly, our choice of metric on this space is the p^{th} Wasserstein distance, and we aggregate nearest neighbours using the associated Wasserstein barycenter. We motivate why the Wasserstein distance is the natural choice for this particular problem in Section 1.2. Accordingly in later sections we also present different numerical setups to demonstrate how to navigate how reactive/robust the algorithm is on different datasets. The latter will depend on the precise application at hand and the modellers appetite for more swift or more robust indicators.

We benchmark our results with two alternative approaches. The first applies the classical k -means algorithm to the first p moments associated to each segment of market returns. The second is a more classical approach: We implement a version of the hidden Markov model (HMM) using appropriate modifications to make results comparable to ours. We test each algorithm both on real and synthetic data. The success of the unsupervised

learning algorithms in the real data setting is evaluated using a marginal maximum mean discrepancy (MMD), a metric arising from a powerful two-sample test which has experienced a notable rise in popularity in recent machine learning literature. Overall, we show that our data-driven and non-parametric methodology accurately partitions financial return series into distinct clusters which are both distinct from each other whilst remaining internally *self-similar*, relative to a more naive approach based on moments or HMMs, which we verify on synthetic parametric data as well as on historical time series, where our algorithm correctly identifies all (now historically known) periods of unusual market activity as they arise.

1.1. The market regime clustering problem. We begin by giving an overview of related literature and existing insights on the MRCP, which is defined as the task of classifying segments of return series $(r_i)_{i \geq 0}$, where

$$r_i = (r_i^1, \dots, r_i^n) \text{ for } n \in \mathbb{N}.$$

Any vector $r_i \in \mathbb{R}^n$ can be associated to an empirical measure $\delta_{r_i} = \frac{1}{n} \sum_{j=1}^n \delta_{r_i^j}$ for $i \geq 0$ with n atoms. Thus, the problem of classifying market regimes is equivalent to assigning labels to probability measures $\mu \in \mathcal{P}_p(\mathbb{R})$, where $\mathcal{P}_p(\mathbb{R})$ is the set of probability measures on \mathbb{R} with finite p^{th} moment.

Historically, approaches to solving the MRCP vary depending on their applications: for instance, a modeller may be interested in regime classification over different time scales: from microseconds, to months or years. Further variations may come according to the modeller's choice of framing the problem; for example, one may wish to segment regimes via change points detection methods, which place (as the name suggests) emphasis on change-points rather than on the nature of the regimes themselves (see Niu, Hao, and Zhang (2016) for an overview). Another example is the more general outlier detection problem, which is a special one-class case of the MRCP, see Cochrane, Foster, Lyons, and Arribas (2020); Kondratyev, Schwarz, and Horvath (2020). Here one is more concerned with identifying anomalous datum as opposed to characterising the distribution $\mu \in \mathcal{P}(\mathbb{R})$ such datum are generated from.

Some of the early attempts at analysing financial return series to extract regime switching signals fall under the umbrella of *technical analysis*, where signals such as temporal moving average crossovers and support level breaches are used as indicators for a regime change for a particular financial asset or a collection of assets, see Achelis (2001) for a comprehensive guide. More rigorous statistical analyses of financial time series have also been employed to detect regime changes. A classical approach is performing dynamic PCA on inter-asset covariance matrices, see Pelletier (2006) for an example on high-dimensional synthetic Markovian asset price paths.

Another classical approach to the MRCP is via Markovian switching models and HMMs. To the author's knowledge, this technique was introduced in the work of Hamilton (1989), in which the state term signifying the regime is given by an AR(1) process. For a more detailed history of the Markov switching model, we refer the reader to Kuan (2002); Lange and Rahbek (2009) and Guidolin (2011). The hidden Markov model approach is not altogether model-free as it makes two main assumptions: first, the latent state variable specifying the current regime is Markovian, and secondly, that the likelihood of observing a return given the latent state variable is given by some parametric distribution, often Gaussian. Other approaches include agent-based models as shown in Lux and Marchesi (2000), Bayesian approaches as seen in Maheu, McCurdy, and Song (2012), or more data-driven approaches as seen in Lahmiri (2016).

Previous work on the problem of clustering families of distributions via non-parametric unsupervised learning approaches has been found in Nielsen, Nock, and Amari (2014), where a modification of the classical k -means algorithm is used to cluster histograms via mixed α -divergences. An approach to empirical distribution clustering via k -means is also given in Henderson, Gallagher, and Eliassi-Rad (2015) in a non-financial context. Other works have utilised the Wasserstein distance for clustering problems, see for instance Li and Wang (2008); Ye, Wu, Wang, and Li (2017), where in the latter distributions are represented as weight-mass pairs, and clustering is considered in the context of images and documents, or Mi, Zhang, Gu, and Wang (2018) for an approach using variational optimal transport. Such approaches are similar to the work in this paper as they often employ classic unsupervised learning algorithms with some modification that allows them to handle distributional datum. Our approach seeks to meld the financial regime-switching and clustering worlds together in an attempt to provide a robust, non-parametric approach to *a posteriori* market regime clustering.

1.2. Motivation for using \mathcal{W}_p . In this section, we motivate our perspective on the clustering problem and the choice of the Wasserstein metric (14) for this purpose, which has recently seen a swift rise in artificial intelligence and machine learning, see for instance Ni, Szpruch, Wiese, Liao, and Xiao (2020).

Given that our clustering problem is defined over the space of probability measures $\mathcal{P}_p(\mathbb{R})$, there exist many candidate one could employ, such as the Kolmogorov-Smirnov (KS)-statistic or the Kullback-Leibler (KL)-divergence. We argue that either of these choices are inadequate for the given problem statement. First, it is well known that the KS-statistic is too tail-sensitive for our requirements, (see Mason and Schuenemeyer (1983)). Secondly, while the KL-divergence has been employed in the literature as a distance function within a clustering algorithm (see Ackermann, Blömer, and Sohler (2008) for a k -medians implementation with generalised Bregman divergences, and Nielsen et al. (2014) for general α -divergences), we note that our clustering problem is over empirical measures, and thus requires an estimation of the probability density function associated to each measure if the KL-divergence is to be employed. Although barycenters with respect to the symmetrized and non-symmetrized versions of the KL-divergence have been shown to exist (see Veldhuis (2002); Nielsen and Nock (2009)), we argue that our approach is much simpler and scalable.

The Wasserstein distance is a natural choice for comparing distributions of points on a metric space (X, d) , as it metrizes the weak convergence¹. Thus, measures that are close in the Wasserstein sense are also close in the classical narrow sense as well.

In our examples we focus on the univariate case $d = 1$. Here, computing the Wasserstein distance in between empirical measures is particularly tractable, though a multivariate characterisation of our results is also possible (see next paragraph). From Proposition 2.5, the algorithm to compute the p -Wasserstein distance between two empirical measures with N atoms is $\mathcal{O}(N \log N)$. It is important to note that other comparative distances share this property. However, the Wasserstein distance also has a natural aggregator candidate in the Wasserstein barycenter, which is fast to calculate in the case where $d = 1$. Other comparative distances either do not have a natural candidate for aggregation (Kolmogorov-Smirnov) or a canonical aggregator which is tractable to calculate (KL-divergence).

¹A sequence $(\mu_n)_{n \geq 1} \subset \mathcal{P}_p(X)$ converges weakly to $\mu \in \mathcal{P}_p(X)$ iff $\mathcal{W}_p(\mu_n, \mu) \rightarrow 0$ as $n \rightarrow \infty$. Moreover, if (X, d) is a Polish space, then $(\mathcal{P}_p(X), \mathcal{W}_p)$ is also Polish (see Ambrosio and Gigli (2013), Theorem 2.6 for the case $p = 2$).

In the case where $d > 1$, the Wasserstein distance is also viable as a choice of metric on $\mathcal{P}_p(\mathbb{R}^d)$. Although \mathcal{W}_p has shown to be an effective tool to tackle the curse of dimensionality associated to sequences on \mathbb{R} , it becomes computationally too demanding to solve when extending to higher-dimensional data (Rabin, Peyré, Delon, and Bernot (2011), Section 2.2). Recently, the sliced Wasserstein distance $\mathcal{W}_{\overline{p}}$ as seen in Rabin et al. (2011), Bonneel, Rabin, Peyré, and Pfister (2015) and Kolouri, Nadjahi, Simsekli, Badeau, and Rohde (2019) has been employed to extend the \mathcal{W}_p to \mathbb{R}^d . It does this by projecting distributions $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$ onto \mathbb{R} via points on the unit sphere in \mathbb{R}^d , and returns the expected one-dimensional Wasserstein distance of such projections. This turns the complex problem of calculating $\mathcal{W}_p(\mu, \nu)$ into an $\mathcal{O}(MN \log N)$ operation, where M is the number of points taken. Recently, Bayraktar and Guo (2021) showed that the max sliced Wasserstein metric $\overline{\mathcal{W}}_p$ is strongly equivalent to the classical Wasserstein distance in the cases $p = 1, 2$.²

1.3. Problem setting and notation. First, we introduce the notion of a data stream.

Definition 1.1 (Set of data streams, Bonnier, Kidger, Arribas, Salvi, and Lyons (2019), Definition 2.1). Let \mathcal{X} be a non-empty set. The set of *streams of data* \mathcal{S} over \mathcal{X} is given by

$$(1) \quad \mathcal{S}(\mathcal{X}) = \{\mathbf{x} = (x_1, \dots, x_n) : x_i \in \mathcal{X}, n \in \mathbb{N}\}.$$

In this paper, we take $\mathcal{X} = \mathbb{R}$ and fix $N \in \mathbb{N}$. In the context of the MRCP, elements $\mathbf{s} = (s_0, \dots, s_N) \in \mathcal{S}(\mathbb{R})$ will be price paths associated to a financial asset.

Given $\mathbf{s} \in \mathcal{S}(\mathbb{R})$, we define the vector of log-returns $r^{\mathbf{s}}$ associated to \mathbf{s} by

$$(2) \quad r_i^{\mathbf{s}} = \log(s_{i+1}) - \log(s_i) \quad \text{for } 0 \leq i \leq N-1,$$

so $r^{\mathbf{s}} \in \mathcal{S}(\mathbb{R})$. We use the following expression to highlight that we may wish partition the original stream of data \mathbf{s} into potentially overlapping segments equal length.

Definition 1.2 (Stream lift, Bonnier et al. (2019), Section 3.3). Let $\mathcal{S}(\mathcal{X})$ be a space of streams over a non-empty set \mathcal{X} . Let \mathcal{V} be another non-empty set, and let $v \geq 1$. We call a function

$$\ell = (\ell^1, \dots, \ell^v) : \mathcal{S}(\mathcal{X}) \rightarrow \mathcal{S}(\mathcal{S}(\mathcal{V}))$$

a *lift* from the space of streams to the space of streams of segments over \mathcal{V} .

Thus, for $\mathbf{s} \in \mathcal{S}(\mathbb{R})$ and $h_1, h_2 \in \mathbb{N}$ with $h_1 > h_2$, we define a lift $\ell := \ell_{h_1, h_2}$ from $\mathcal{S}(\mathbb{R})$ to $\mathcal{S}(\mathcal{S}(\mathbb{R}))$ via

$$(3) \quad \ell^i(\mathbf{s}) = (s_{1+h_2(i-1)}, \dots, s_{1+h_1+h_2(i-1)}) \quad \text{for } i = 1, \dots, M,$$

where $M := 1 + \lfloor \frac{N-h_2}{h_2} \rfloor$ is the maximum number of partitions with length h_1 that can be extracted from $\mathbf{s} \in \mathcal{S}(\mathbb{R})$ with sliding window offset parameter h_2 . We obtain the stream of segments by applying ℓ to $r^{\mathbf{s}}$. Finally, as we are interested in clustering over a compact set of probability measures, we define the following.

Definition 1.3 (Empirical measure, Goodfellow, Bengio, and Courville (2016), Section 3.9.5). Let $\mathbf{x} \in \mathcal{S}(\mathbb{R})$ such that $\mathbf{x} = (x_1, \dots, x_N)$ for $N \in \mathbb{N}$. Furthermore, let

$$Q^j : \mathcal{S}(\mathbb{R}) \rightarrow \mathbb{R}$$

²The authors would like to thank Claude Martini and Frédéric Patras for informing us of this.

be the function which extracts the j^{th} order statistic of \mathbf{x} , for $j = 1, \dots, N$. Then, the cumulative distribution function of the *empirical measure* $\mu \in \mathcal{P}_p(\mathbb{R})$ associated to \mathbf{x} is defined as

$$(4) \quad \mu^{\mathbf{x}}((-\infty, x]) = \frac{1}{N} \sum_{i=1}^N \chi_{\{Q^i(\mathbf{x}) \leq x\}}(x),$$

where $\chi : \mathbb{R} \rightarrow [0, 1]$ is the indicator function.

Thus, we can associate to each segment of data $r_i \in \ell(r^s)$ the empirical measure μ_i for $i = 1, \dots, M$. This gives us a family of measures

$$(5) \quad \mathcal{K} = \{(\mu_1, \dots, \mu_M) : \mu_i \in \mathcal{P}_p(\mathbb{R}) \text{ for } i = 1, \dots, M\}.$$

It is this family \mathcal{K} which will be the subject of our clustering algorithm.

1.4. The k -means algorithm. Suppose $X = \{(\mathbf{x}_1, \dots, \mathbf{x}_N) : \mathbf{x}_i \in V\} \in \mathcal{S}(V)$ is a stream of data over a normed vector space $(V, \|\cdot\|_V)$. We further assume that each $\mathbf{x}_i = (x_1^i, \dots, x_d^i)$ has been standardised coordinate-wise, that is,

$$(6) \quad \mathbb{E}[\{x_j^i\}_{1 \leq i \leq N}] = 0 \text{ and } \text{Var}(\{x_j^i\}_{1 \leq i \leq N}) = 1 \quad \text{for } j = 1, \dots, d.$$

The *k-means clustering algorithm* is a unsupervised vector quantization method which assigns elements of X to k distinct clusters. Each of these clusters are defined by central elements $\bar{\mathbf{x}} := \{\bar{\mathbf{x}}_j\}_{j=1, \dots, k}$ called *centroids*, which are often initially sampled from X .

At each step $n \in \mathbb{N}$ of the algorithm, one first calculates the *nearest neighbours*

$$(7) \quad \mathcal{C}_l^n := \left\{ \mathbf{x}_i \in X : \arg \min_{j=1, \dots, k} d(\mathbf{x}_i, \bar{\mathbf{x}}_j^{n-1}) = l \right\}$$

associated to each $\bar{\mathbf{x}}_j^{n-1}$ for $j = 1, \dots, k$, where $d : V \times V \rightarrow [0, +\infty)$ is the metric induced by the norm on V .

Remark 1.4. Classically, one chooses $(V, \|\cdot\|_V) = (\mathbb{R}^d, \|\cdot\|_{\mathbb{R}^d})$, but we note here that any normed vector space could be chosen.

Each set \mathcal{C}_l^n is then aggregated into a new centroid \mathbf{x}_l^n for $l = 1, \dots, k$ via a function $\alpha : 2^V \rightarrow V$, so

$$\bar{\mathbf{x}}_l^n := \alpha(\mathcal{C}_l^n) \quad \text{for } l = 1, \dots, k.$$

For a given a tolerance level $\varepsilon > 0$ and a loss function $l : V^k \times V^k \rightarrow [0, +\infty)$, the k -means algorithm terminates at step $n \in \mathbb{N}$ if the stopping condition

$$l(\bar{\mathbf{x}}^n, \bar{\mathbf{x}}^{n-1}) < \varepsilon$$

is satisfied. The algorithm outputs the final clusters $\mathcal{C}^* = \{\mathcal{C}_l^n\}_{l=1, \dots, k}$ and the k quantizations $\bar{\mathbf{x}}^n = \{\bar{\mathbf{x}}_l^n\}_{l=1, \dots, k}$. We conclude this section with the assumptions associated to the k -means algorithm that, if satisfied, will result in uniform and isotropic clustering of a data set X .

Proposition 1.5 (Kanungo et al. (2002)). *Given data X , the k -means algorithm produces k suitable clusters if the following is true:*

1. *There exist k natural clusters in the data X .*
2. *Each cluster within X is of roughly equal size.*

3. *Within-cluster variation (cf. Definition A.2) is uniform. That is, for $\delta_2 > 0$ small we have that*

$$|\text{WC}(\mathcal{C}_i) - \text{WC}(\mathcal{C}_j)| < \delta_2 \quad \text{for } i, j = 1, \dots, k \text{ and } i \neq j,$$

4. *Clusters are spherical in shape, so we expect the nearest neighbours \mathcal{C}_j to the j^{th} centroid \bar{x}_j to be contained within a ball $B(\bar{x}_j, \delta)$ where $\delta > 0$ is uniform across all clusters $j = 1, \dots, k$.*

If conditions (1)-(4) are satisfied, then optimal clusterings \mathcal{C}^* will be suitable.

Counterexamples to suitability include forcing k clusters on data with fewer than k natural clusters available. Another classical example is the problem of clustering concentric data $X \subset \mathbb{R}^2$, which violates assumption (4) in Proposition 1.5. We note that do exist other clustering algorithms which do not share the drawbacks of k -means, in that they do not make assumptions regarding the number or shape of the clusters (hierarchical clustering), nor do they enforce that data points x_i belong to one individual cluster (fuzzy c-means clustering, see for instance Cannon, Dave, and Bezdek (1986)). Exploring other clustering algorithms for the MRCP is a topic for future research.

1.5. The maximum mean discrepancy. Typically, the evaluation of derived clusters from a data stream $\mathcal{S}(\mathcal{X})$ is done by assessing the final total cluster variation $\text{TC}(\mathcal{C}^*)$, also referred to as *inertia* (see Definition A.3). Here, \mathcal{C}^* are the final clusters obtained from a given run of the k -means algorithm. Naturally the value of $\text{TC}(\mathcal{C}^*)$ is dependent on the normed vector space $(V_1, \|\cdot\|_{V_1})$ one decides to cluster the stream of data \mathcal{X} over. Of course one could make a different choice $(V_2, \|\cdot\|_{V_2})$ by transforming sets of datum $A \subset \mathcal{S}(\mathcal{X})$, see Section 3.1. Since the total cluster variation depends on V , one cannot use it in evaluation between clusterings on different choices of V .

In this section, we outline an integrable probability metric on the space of distributions called the *maximum mean discrepancy (MMD)*, which will be used as part of a robust methodology for confirming goodness-of-fit of clustered market regimes. The MMD has been shown to be a robust estimator under both dependence and presence of outliers and has been employed frequently in the quantitative finance and machine learning literature, see Buehler et al. (2020); Alquier, Chérif-Abdellatif, Derumigny, and Fermanian (2020); Chérif-Abdellatif and Alquier (2021), and Briol, Barp, Duncan, and Girolami (2019).

We provide a brief introduction here and refer the reader to Gretton, Borgwardt, Rasch, Schölkopf, and Smola (2012); Gretton, Borgwardt, Rasch, Schölkopf, and Smola (2007); Gretton, Fukumizu, Harchaoui, and Sriperumbudur (2009) for further details. We begin by introducing the following.

Problem 1.6 (Two-sample test, Gretton et al. (2012), Problem 1). Let (\mathcal{X}, d) be a metric space. Suppose X and Y are independent random variables defined on \mathcal{X} . Suppose that $X_{\#}\mathbb{P} = \mu$ and $Y_{\#}\mathbb{P} = \nu$, where $\mu, \nu \in \mathcal{P}(\mathcal{X})$ are Borel. If we draw samples $x = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_m)$ where $x_i \sim \mu$ for $i = 1, \dots, n$ and $y_j \sim \nu$ for $j = 1, \dots, m$, when can we determine if $\mu \neq \nu$? That is, we wish to implement a test for the *two-sample problem*

$$(8) \quad H_0 : \mu = \nu \text{ against } H_1 : \mu \neq \nu.$$

We introduce the following test statistic associated to Problem 1.6.

Definition 1.7 (Maximum mean discrepancy, Gretton et al. (2012), Definition 2). Let \mathcal{F} be a class of functions $f : \mathcal{X} \rightarrow \mathbb{R}$ and let μ, ν be defined as in Problem 1.6. Then, the *maximum mean discrepancy* (MMD) between μ and ν is defined as

$$(9) \quad \text{MMD}[\mathcal{F}, \mu, \nu] := \sup_{f \in \mathcal{F}} \left(\mathbb{E}_\mu[f(x)] - \mathbb{E}_\nu[f(y)] \right).$$

If $x = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_m)$ are samples where $x_i \sim \mu$ and $y_j \sim \nu$, then a *biased empirical estimate* of the MMD is given by

$$(10) \quad \text{MMD}_b[\mathcal{F}, x, y] := \sup_{f \in \mathcal{F}} \left[\frac{1}{n} \sum_{i=1}^n f(x_i) - \frac{1}{m} \sum_{j=1}^m f(y_j) \right]$$

Clearly, the value of the MMD between two measures is determined by the function class \mathcal{F} one decides to calculate the supremum in eq. (9) over. In particular, it is not even guaranteed to be a metric. Often the MMD is employed in the context of studying mean differences between datum in a typically higher-dimensional feature space. This motivates the use of kernel methods to define \mathcal{F} , which is often chosen to be the unit ball in a reproducing kernel Hilbert space (RKHS) (\mathcal{H}, κ) , where $\kappa : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is the associated reproducing kernel. For compact \mathcal{X} , *universality* of the associated kernel κ (cf. Definition B.3) is a sufficient condition for making the associated MMD a metric. If \mathcal{X} is non-compact, the following (related) property is enough to guarantee that this is the case.

Definition 1.8 (Characteristic kernel, Fukumizu, Gretton, Schölkopf, and Sriperumbudur (2009), Section 2). Let \mathcal{X} be a non-empty set. A kernel κ on \mathcal{X} is called *characteristic* if the mean mapping

$$(11) \quad \mu \mapsto \mathbb{E}_{X \sim \mu}[\kappa(\cdot, X)]$$

is injective.

The *Gaussian kernel*

$$(12) \quad \kappa_G : \mathbb{R}^d \times \mathbb{R}^d \rightarrow [0, +\infty), \quad \kappa_G(x, y) = \exp(-\|x - y\|_{\mathbb{R}^d}^2 / 2\sigma^2)$$

is characteristic to the set of Borel measures on \mathcal{X} and indeed makes the MMD a metric on $\mathcal{P}(\mathcal{X})$. For more details we refer the reader to Theorem 2 in Fukumizu, Gretton, Sun, and Schölkopf (2007) or to Appendix B, Theorem B.4 for more details.

We will use the MMD with $\mathcal{F} = (\mathcal{H}, \kappa_G)$ to validate how effective a given clustering algorithm is in the case that we are unable to infer true regime labels, i.e., when we are working with real data. A key notion to define how similar a collection of samples are to each other is the following.

Definition 1.9 (Within-cluster self-similarity (homogeneity)). Let $X \in \mathcal{S}(\mathcal{X})$ be a stream of data with N observations. Let \mathcal{F} be the unit ball in a universal RKHS \mathcal{H} . For $n, m \in \mathbb{N}$, we define the *self-similarity score* associated to \mathcal{X} to be

$$(13) \quad \text{Sim}(X) = \text{Median} \left((\text{MMD}_b^2[\mathcal{F}, x_i, y_i])_{1 \leq i \leq n} \right),$$

where $x_i = (x_1^i, \dots, x_m^i)$ and $y_i = (y_1^i, \dots, y_m^i)$ are samples drawn pairwise from \mathcal{X} for $i = 1, \dots, n$.

Remark 1.10. The *true self-similarity score* is obtained by calculating the biased MMD (10) for all unique combinations of pairwise samples. For computational reasons, we often calculate eq.(13) from $n \ll \binom{N}{2}$ iterations.

2. k -MEANS ON THE SPACE OF DISTRIBUTIONS

In this section, we outline our modification to the k -means algorithm which allows us to cluster the set (5) directly on the space of probability measures with finite p^{th} moment. Central to this paper is the following distance metric on $\mathcal{P}_p(\mathbb{R})$.

Definition 2.1 (p -Wasserstein distance, Ambrosio, Gigli, and Savare (2005)). Suppose (X, d) is a separable Radon space. The p -th Wasserstein distance between measures $\mu, \nu \in \mathcal{P}_p(X)$ is defined by

$$(14) \quad \mathcal{W}_p^p(\mu, \nu) := \min_{\mathbb{P} \in \Pi(\mu, \nu)} \left\{ \int_{X \times X} d(x, y)^p \mathbb{P}(dx, dy) \right\},$$

where

$$\Pi(\mu, \nu) := \{ \mathbb{P} \in \mathcal{P}(X \times X) : \mathbb{P}(A \times X) = \mu(A), \mathbb{P}(X \times B) = \nu(B) \}$$

is the set of *transport plans* between μ and ν .

The p -Wasserstein distance (14) is the solution to the Kantorovich-type optimal transportation problem between measures μ and ν for the cost function $c(x, y) = d(x, y)^p$. For our applications, existence of an optimal plan $\mathbb{P}^* \in \Pi(\mu, \nu)$ realising the Wasserstein distance between measures $\mu, \nu \in \mathcal{P}(\mathbb{R})$ is guaranteed by continuity of the metric $d(x, y) = |x - y|^p$ and the fact that μ, ν are compactly supported. We refer the reader to Santambrogio (2015) for further details.

Remark 2.2 (Relationship to the MMD). The Wasserstein distance (14), via its equivalent dual formulation, is a special case of an integral probability metric (see, for instance, Wang, Gao, and Xie (2021), Definition 1). In the case where $p = 1$, the dual representation is given by

$$(15) \quad \mathcal{W}_1(\mu, \nu) = \sup_{f \in \text{Lip}_1(X)} \left\{ \int_X f d(\mu - \nu) \right\},$$

where $\text{Lip}_1(X)$ denotes the space of continuous \mathbb{R} -functions over X with Lipschitz constant $L \leq 1$. Since μ, ν are probability measures, we can write (15) as

$$\mathcal{W}_1(\mu, \nu) = \sup_{\|f\|_{\text{Lip}_1} \leq 1} \{ \mathbb{E}_\mu[f(x)] - \mathbb{E}_\nu[f(y)] \}.$$

Thus $\mathcal{W}_1(\mu, \nu)$ is an integrable probability metric over the function class \mathcal{F} , which is given by the unit ball in the space of functions

$$\text{Lip}(X) = \{ f : X \rightarrow \mathbb{R} : f \text{ continuous, } \|f\|_{\text{Lip}} < +\infty \},$$

where

$$\|f\|_{\text{Lip}} = \sup_{x \neq y} \frac{|f(x) - f(y)|}{d(x, y)}.$$

In what follows, we choose the p -Wasserstein distance to be our metric on $\mathcal{P}_p(\mathbb{R})$. Our next decision we need to make is how we aggregate nearest neighbours \mathcal{C}_l into central elements $\bar{\mu}_l \in \mathcal{P}_p(\mathbb{R})$ for $l = 1, \dots, k$. One of the advantages of choosing the Wasserstein distance \mathcal{W}_p to be the metric we apply on our clustering space is the existence of the following, which gives a natural way to “average” a family of measures under \mathcal{W}_p .

Definition 2.3 (Wasserstein barycenter). Suppose (X, d) is a separable Radon space and let $\mathcal{K} = \{\mu_i\}_{i \geq 1} \subset \mathcal{P}(X)$ be a family of Radon measures. Define the p -Wasserstein barycenter $\bar{\mu}$ of \mathcal{K} to be

$$(16) \quad \bar{\mu} = \arg \min_{\nu \in \mathcal{P}(X)} \sum_{\mu_i \in \mathcal{K}} \mathcal{W}_p(\mu_i, \nu).$$

Remark 2.4. If $\{\mu_i\}_{i \geq 1}$ are a family of measures associated to a cluster $\mathcal{C}_l, l = 1, \dots, k$, then the Wasserstein barycenter (16) is the measure $\bar{\mu} \in \mathcal{P}_p(\mathbb{R})$ which minimises the within-cluster variation $\text{WC}(\mathcal{C}_l)$ from Definition A.2.

Since the k -means algorithm requires repeated evaluations of elements on clustering space under the given metric, tractability of the optimisation from eq. (14) is relevant. In the case where measures $\mu, \nu \in \mathcal{P}(\mathbb{R})$ are absolutely continuous, there exist a closed-form solution to (14).

Proposition 2.5 (Kolouri et al. (2019), Equation (3)). Suppose $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$ and let $d = 1$. Moreover, suppose that μ, ν are absolutely continuous with respect to the Lebesgue measure on \mathbb{R} . Then, the p -Wasserstein distance $\mathcal{W}_1(\mu, \nu)$ is given by

$$(17) \quad \mathcal{W}_p(\mu, \nu) = \left(\int_0^1 |F_\mu^{-1}(z) - F_\nu^{-1}(z)|^p dz \right)^{1/p},$$

where the quantile function $F_\mu^{-1} : [0, 1) \rightarrow \mathbb{R}$ is defined as

$$(18) \quad F_\mu^{-1}(z) = \inf\{x : F_\mu(x) > z\}.$$

Proof. A consequence of the fact that the (unique) optimal transport map pushing μ onto ν is given by $T(x) = (F_\nu^{-1} \circ F_\mu)(x)$, and applying a change of variables. \square

In what follows, we assume that μ, ν are empirical measures with equal numbers of atoms $N \in \mathbb{N}$ (this will be the case in our experimental setup). Recalling Definition 1.3, we may write them as

$$(19) \quad \mu((-\infty, x]) = \frac{1}{N} \sum_{i=1}^N \chi_{\alpha_i \leq x}(x), \quad \nu((-\infty, x]) = \frac{1}{N} \sum_{i=1}^N \chi_{\beta_i \leq x}(x)$$

where $(\alpha_i)_{1 \leq i \leq N}$ and $(\beta_i)_{1 \leq i \leq N}$ are increasing sequences corresponding to the atoms of μ and ν . We wish to use (17) to obtain a closed-form expression for the Wasserstein distance between the two measures. Thus, we must consider the well-posedness of (18): every empirical measure on \mathbb{R} is Radon, and thus one can associate to μ, ν a right-continuous function of finite variation $A_t : \mathbb{R} \rightarrow [0, 1]$ given by $A_t = \mu((-\infty, t))$ (Revuz and Yor (2004), Theorem 4.3). The function A_t possesses a right-continuous inverse which is nothing but the quantile function from (18), which (in the case of μ) can be written as

$$(20) \quad F_\mu^{-1}(z) = \alpha_i \quad \text{for all } z \in \left[\frac{i-1}{N}, \frac{i}{N} \right), \quad i = 1, \dots, N.$$

Moreover $F_\mu^{-1}(z) = 0$ for all $z < \alpha_1$. Applying (20) to (17), the Wasserstein distance between the empirical measures μ and ν is given by

$$\begin{aligned} \mathcal{W}_p(\mu, \nu)^p &= \sum_{i=1}^N \int_{\frac{i-1}{N}}^{\frac{i}{N}} |F_\mu^{-1}(z) - F_\nu^{-1}(z)|^p dz \\ (21) \quad &= \frac{1}{N} \sum_{i=1}^N |\alpha_i - \beta_i|^p. \end{aligned}$$

Thus, calculating the Wasserstein distance between two empirical measures is $\mathcal{O}(N)$, assuming the atoms of each measure are already sorted ascending, and N is the number of atoms. If not, it is an $\mathcal{O}(N \log N)$ operation. This representation also makes calculating the Wasserstein barycenter (16) simple in the case where $p = 1$ and some assumptions are made on the number of atoms present in each measure.

Proposition 2.6 (Wasserstein barycenter, empirical measures). *Suppose that $\{\mu_i\}_{1 \leq i \leq M}$ are a family of empirical probability measures, each with N atoms $(\alpha_j^i)_{1 \leq j \leq N} \subset \mathbb{R}^N$. Let*

$$a_j = \text{Median}(\alpha_j^1, \dots, \alpha_j^M) \quad \text{for } j = 1, \dots, N.$$

Then, the cumulative distribution function of the Wasserstein barycenter $\bar{\mu} \in \mathcal{P}_1(\mathbb{R})$ over $\{\mu_i\}_{1 \leq i \leq M}$ with respect to the 1-Wasserstein distance is given by

$$(22) \quad \bar{\mu}((-\infty, x]) = \frac{1}{N} \sum_{i=1}^N \chi_{a_i \leq x}(x).$$

Moreover, $\bar{\mu}$ is not necessarily unique.

Proof. See Appendix C.1. □

The last specification we need to make is regarding the loss function. We do this in the natural way by replacing the squared Euclidean distance in standard k -means by the p -Wasserstein distance. Let $\bar{\mu}^n = (\bar{\mu}_i^n)_{1 \leq i \leq k}$ be the centroids obtained after step n of the Wasserstein k -means algorithm. Therefore, our loss function $l : \mathcal{P}_p(\mathbb{R})^k \times \mathcal{P}_p(\mathbb{R})^k \rightarrow [0, +\infty)$ is given by

$$(23) \quad l(\bar{\mu}^{n-1}, \bar{\mu}^n) = \sum_{i=1}^k \mathcal{W}_p(\bar{\mu}_i^{n-1}, \bar{\mu}_i^n).$$

Our stopping rule is unchanged; that is, for a given $\varepsilon > 0$ we terminate the algorithm at step n if $l(\bar{\mu}^{n-1}, \bar{\mu}^n) < \varepsilon$. We give a full statement of the algorithm with the following.

Definition 2.7 (WK-means algorithm). Let $\mathcal{K} \subset \mathcal{P}_p(\mathbb{R})$ be a family of measures with finite p^{th} moment. We refer to the k -means clustering algorithm on $(\mathcal{P}_p(\mathbb{R}), \mathcal{W}_p)$, with aggregation method given by the Wasserstein barycenter from Definition 2.3 and loss function given by (23) as the *Wasserstein k -means algorithm*, or *WK-means*.

The explicit algorithm can be found in Appendix C.

3. METHODOLOGY AND NUMERICAL RESULTS

In this section, we cover the methods used to test the WK-means algorithm on stock data. Initially, we test both algorithms on real data. Validation of each clustering algorithm was conducted using the MMD test statistic (10). Finally, we tested both algorithms on

synthetic data generated via two different models: one where the associated log-returns were distributed normally, and another where they were not.

3.1. Alternative clustering algorithms as benchmarks. In this section we introduce two methods we use to benchmark our approach.

3.1.1. k -means with statistical moments. A natural and more classical approach to clustering regimes may involve studying the first $p \in \mathbb{N}$ raw moments associated to each measure $\mu \in \mathcal{K}$. With this in mind, consider the image of \mathcal{K} from (5) under the function

$$(24) \quad \varphi^p(\mu) = \left(\frac{1}{n!} \int_{\mathbb{R}} x^n \mu(dx) \right)_{1 \leq n \leq p},$$

which is the *truncated* unstandardised p^{th} -moment map. As each $\mu \in \mathcal{K}$ is a sum of Dirac masses, each element of $\varphi^p(\mu)$ is finite. Thus, for a given $p > 1$ we obtain

$$(25) \quad \varphi^p(\mathcal{K}) = \{(\varphi^p(\mu_1), \dots, \varphi^p(\mu_M)) : \varphi^p(\mu_i) \in \mathbb{R}^p \text{ for } i = 1, \dots, M\}.$$

After standardising each element of $\varphi^p(\mathcal{K})$ component-wise (cf. Remark 3.2). This motivates the following definition.

Definition 3.1 (Moment k -means). Let $\mathcal{K} \subset \mathcal{P}_p(\mathbb{R})$ be a family of measures. For $p \geq 1$, associate to each $\mu_i \in \mathcal{K}$ the \mathbb{R}^p -vector $\varphi^p(\mu_i)$ for $i = 1, \dots, M$, where $\varphi^p : \mathcal{P}_p(\mathbb{R}) \rightarrow \mathbb{R}^p$ is the p -moment map from (24).

Then, *moment k -means algorithm*, or *MK-means*, is given by applying Algorithm 1 to the stream of data $\varphi^p(\mathcal{K})$ from (25). See Appendix A for more details.

Remark 3.2 (Magnitude of moments). The function φ^p defined in (24) outputs the first p raw moments associated to a measure $\mu \in \mathcal{P}_p(\mathbb{R})$. Often, moments that appear earlier in the sequence $(\varphi^p(\mu)_i)_{1 \leq i \leq p}$ will be of significantly larger magnitude than those that appear later. In order for the k -means algorithm to not place undue emphasis on these moments, it is critical that each slice $\{\varphi_p(\mu_i)_j\}_{1 \leq i \leq M}$ is standardised according to equation (6) for $1 \leq j \leq p$.

3.1.2. Hidden Markov model. As mentioned in Subsection 1.1, a more classical approach to market regime clustering involves fitting a *hidden Markov model (HMM)* to observed time series data $\mathbf{x} \in \mathcal{S}(\mathbb{R})$. Much like classical k -means, here one assumes that there exist $k \in \mathbb{N}$ hidden latent states $\{1, \dots, k\}$ which govern the dynamics of \mathbf{x} . The transition between the latent states is assumed Markovian, and although such states are not directly observable they are represented by a transition density $f(x | z_l, \theta_l)$ where z_l is the given latent state and θ_l are parameters associated to the state, for $l = 1, \dots, k$. The most common choice of likelihood function is a Gaussian one. We refer the reader to Dias, Vermunt, and Ramos (2015) for more details.

As another point of comparison to our approach, we fit a Gaussian HMM in both our real and synthetic data experiments. A main point of difference here is that the HMM does not cluster sets of returns: instead, it associates returns at time t to a given latent state. We thus can derive accuracy statistics in the case where we run the HMM over synthetic data, but for real data our validation method using the MMD is not possible.

3.2. Validation on real data. In this section, we give results from each algorithm on real data.

3.2.1. Data and hyperparameters. We begin by testing both algorithms on market data. In particular, we use one-hourly log-returns $r^s \in \mathcal{S}(\mathbb{R})$ associated to SPY from 2005-01-03 to 2020-12-31. Recalling Definition 1.2 and (3), we set the hyper-parameters $(h_1, h_2) = (35, 28)$. This roughly partitions the time-series into weeks, with adjacent partitions within one day of each other. We defer discussions regarding choices of hyperparameters to Section 3.4.

Regarding the number of clusters, we set $k = 2$, reflecting an aforementioned stylised fact regarding financial markets (the existence of bull and bear cycles on a macro scale). We note that other values of k do have financial interpretations - for examples, see Maheu et al. (2012).

We ran the two algorithms over the lifted stream of data $\ell(r^s)$. For each algorithm, we obtained centroids $\{\bar{\mu}_i\}_{i=1,2}$ and sets $\{\mathcal{C}_l\}_{l=1,2}$ with $\mathcal{C}_l \subset \mathcal{P}_p(\mathbb{R})$ being the nearest neighbours corresponding to the l^{th} centroid. In order to display our results, we use two types of plot. The first is the projection of each distribution $\mu \in \mathcal{K}$ onto \mathbb{R}^2 via the map

$$f_p : \mathcal{P}_p(\mathbb{R}) \rightarrow \mathbb{R}^2, \\ \mu \mapsto \left(\sqrt{\text{Var}(\mu)}, \mathbb{E}[\mu] \right),$$

that is, a scatter plot of each measure in mean-variance space. We colour these points according to their centroid membership. Points coloured green correspond to the cluster with lower variance than those coloured red. The second plot we make use of is a time-series plot of stock values S , where each partition has been coloured corresponding to its centroid membership. Given that the empirical distributions overlap, a given timestamp may be classified into multiple clusters. Thus, we colour these points according to their average centroid memberships. In the case of $k = 2$ with the hyperparameters $(h_1, h_2) = (35, 28)$, a single return r_i^s can potentially belong to 5 different empirical measures. Thus, there are potentially $\binom{5}{2=6}$ total centroid membership combinations associated to r_i^s . We colour these sections of the price path accordingly.

3.2.2. Real data results. We first present the results of running either algorithm over hourly SPY data. Figure 1 gives the scatter plots of each empirical measure $\mu \in \mathcal{K}$ coloured according to its cluster membership. The centroids are marked by crosses and coloured accordingly.



Figure 1. Plots of MK-means and WK-means clusters in mean-variance space.

From Figure 1b, we see that the WK-means classifications are much less susceptible to outlier distributions than the MK-means algorithm. We also note that the Wasserstein

approach demarcates distributions $\mu \in \mathcal{K}$ by variance, which one naturally expects in a financial market setting. Although Figure 1a appears to do the same, it is hard to state this definitively as the clustering algorithm seems to primarily group outlier distributions.

This is made apparent in the graphs presented in Figure 2, where we have associated to each $\mu \in \mathcal{K}$ the partition of $\mathbf{s} \in \mathcal{S}(\mathbb{R})$ it is generated from.

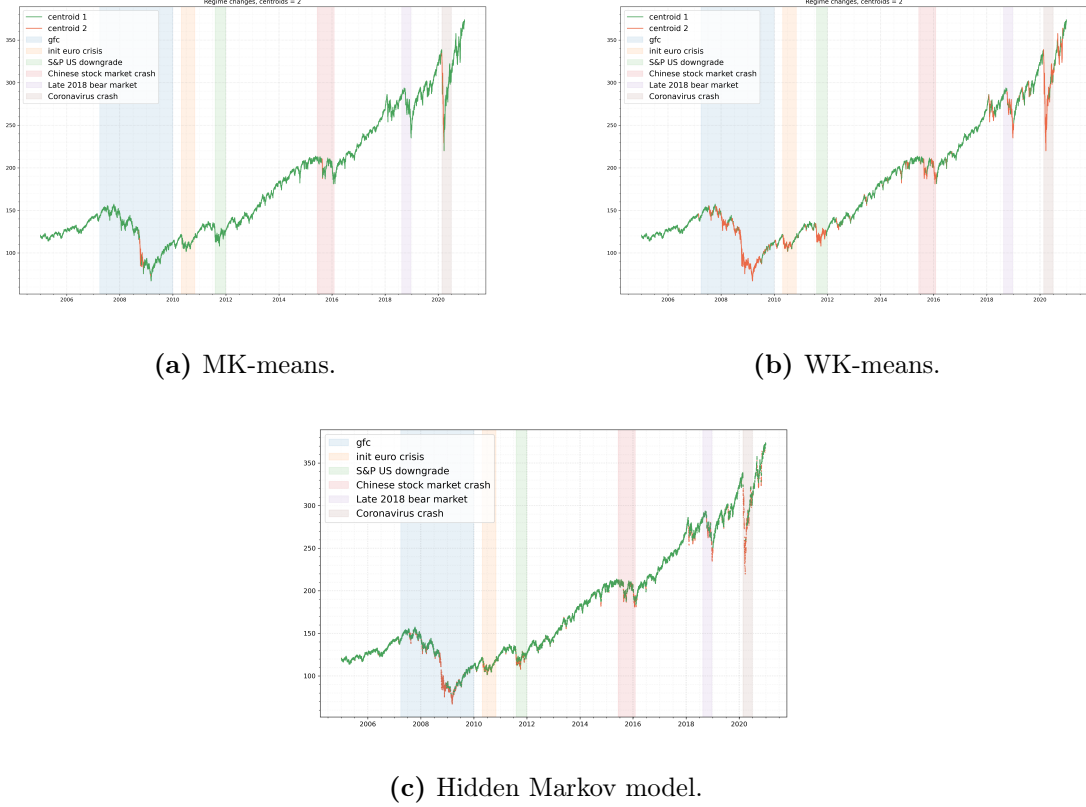


Figure 2. Historical cluster colouring on SPY price path.

Figures 2a and 2b show that both algorithms are able to separately classify periods of returns associated to the global financial crisis and the more recent market instability due to the coronavirus pandemic. However, only the WK-means algorithm is able to distinguish between more subtle periods of stock market volatility: the beginning of the Eurozone/Greek debt crisis in 2010, the S&P US credit-rating downgrade in 2011, and the 2015/16 Chinese stock market crash were tagged as periods of regime change only by the WK-means algorithm. Note that we can include an example of a Hidden Markov model run with this plot, in Figure 2c. We note that the results from this approach seem to sit somewhere between the Wasserstein and moment algorithms.

3.2.3. Validation methods. In order to compare clusterings obtained from both algorithms, we use the marginal MMD test introduced in Section 1.5. The two methods of evaluation we consider are applied both between and within clusters $\mathcal{C} = \{\mathcal{C}_l\}_{l=1,2}$. Moreover, both methods of evaluation are similar, in that they involve bootstrapping the distribution of MMD_b^2 between two sets of samples.

More generally, the definition of an optimal clustering over a set of data is not well defined. Heuristically, we would like individual clusters to contain objects that are similar to each other whilst being distinct from objects in other clusters. We note that there do already

exist several indexes used to evaluate the result of a given k -means clustering, which we recall here.

Definition 3.3 (Davies-Bouldin index, Davies and Bouldin (1979)). Suppose $(V, \|\cdot\|_V)$ is a normed vector space. Let $\{(\bar{x}_l, \mathcal{C}_l)\}_{l=1}^k$ be k centroids and clusters over $X \in \mathcal{S}(\mathcal{X})$, obtained by applying the k -means algorithm characterised by the functions

$$\begin{aligned}\varphi : \mathcal{X} &\rightarrow V, \\ a : 2^V &\rightarrow V, \text{ and} \\ l : V^K \times V^k &\rightarrow [0, +\infty).\end{aligned}$$

Suppose $d : V \times V \rightarrow [0, +\infty)$ is the metric induced by the norm on V . Let

$$d_l = \frac{1}{|\mathcal{C}_l|} \sum_{x \in \mathcal{C}_l} d(x, \bar{x}_l)$$

be the average distance of cluster elements $x \in \mathcal{C}_l$ to the central element \bar{x}_l for $l = 1, \dots, k$. Then, the *Davies-Bouldin index* is given by

$$(26) \quad DB \left(\{(\bar{x}_l, \mathcal{C}_l)\}_{l=1}^k \right) = \frac{1}{k} \sum_{i=1}^k \max_{j \neq i} \frac{d_i + d_j}{d(\bar{x}_i, \bar{x}_j)}.$$

Lower values of (26) are indicative of a better clustering.

Definition 3.4 (Dunn index, Dunn (1974)). With the same notation as Definition 3.3, define

$$\underline{d}_{ij} = \min_{x \in \mathcal{C}_i, y \in \mathcal{C}_j} d(x, y)$$

to be the smallest distance between elements of each cluster. Also define

$$\bar{d}_l = \max_{x, y \in \mathcal{C}_l} d(x, y)$$

to be the largest intra-cluster distance between all clusters $\{\mathcal{C}_l\}_{1 \leq l \leq k}$. Then, the *Dunn index* is given by

$$(27) \quad D \left(\{(\bar{x}_l, \mathcal{C}_l)\}_{l=1}^k \right) = \frac{\min_{1 \leq i, j \leq k} \underline{d}_{ij}}{\max_{1 \leq l \leq k} \bar{d}_l}.$$

Larger values of (3.4) are indicative of a better clustering.

Definition 3.5 (Silhouette coefficient, Rousseeuw (1987)). With the same notation as Definition 3.3, define

$$b_i = \min_{i \neq j} \frac{1}{|\mathcal{C}_j|} \sum_{y \in \mathcal{C}_j} d(x_i, y),$$

and

$$a_i = \frac{1}{|\mathcal{C}_i|} \sum_{y \in \mathcal{C}_i} d(x_i, y).$$

for any $x_i \in \mathcal{C}_l$, $l = 1, \dots, k$. Then, the *Silhouette coefficient* of the point x_i is given by

$$(28) \quad S(i) = \frac{b_i - a_i}{\max(a_i, b_i)}.$$

From (28), we can see that $-1 \leq S(i) \leq 1$. Higher values of $S(i)$ mean that the point x_i was appropriately allocated to cluster \mathcal{C}_l .

Algorithm	Davies-Bouldin	Dunn	\bar{S}_α
Moment	0.8604	9.2×10^{-3}	0.8008
Wasserstein	1.1075	7.3×10^{-3}	0.5093

Table 1. Scores for MK- and WK-means algorithms using traditional k -means index evaluation methods, typical run.

Remark 3.6. It is often computationally expensive to calculate the (28) for every point \mathbf{x} . Thus, we often use an estimate from fewer samples.

For $0 < \alpha \leq 1$, define $\lambda_l = \lfloor \alpha |\mathcal{C}_l| \rfloor$ and let $(n_k^l)_{k=1}^{\lambda_l}$ be an increasing sub-sequence of $\{1, \dots, |\mathcal{C}_l|\}$, for $l = 1, \dots, k$. Then, the α -average *Silhouette coefficient* \bar{S}_α is given by

$$(29) \quad \bar{S}_\alpha = \frac{1}{k} \sum_{l=1}^k \left(\frac{1}{\lambda_l} \sum_{k=1}^{\lambda_l} S(n_k^l) \right).$$

Definitions 3.3, 3.4 and 3.5 are often used to evaluate clusters derived from a standard k -means algorithm for different values of k . We will see that using them to compare clusterings between the MK- and WK-means approaches (for the same value of k) does not capture how appropriate clusterings are in reference to the MRCP. Firstly, such indexes are not agnostic to the choice of $(V, \|\cdot\|_V)$ and are thus not comparable between algorithms. Secondly, a more appropriate validation method for the MRCP would be between regimes $\mu \in \mathcal{S}(\mathcal{P}_p(\mathbb{R}))$, as opposed to elements of the clustering space $\varphi(\mu) \in V$. Thus, as an integrable probability metric, the MMD is a more suitable choice to be used to evaluate the appropriateness of either clustering algorithm. Nevertheless we will report the values of these metrics for completeness.

For our between-cluster evaluation, we proceed as follows. Given sets $\mathcal{C}_1, \mathcal{C}_2$ obtained via either the moment- or WK-means clustering algorithm, draw $n \in \mathbb{N}$ pairwise samples $(\mu_i, \nu_i) \in \mathcal{C}_1 \times \mathcal{C}_2$ for $i = 1, \dots, n$. We represent each empirical measure $\mu_i, \nu_i \in \mathcal{P}_p(\mathbb{R})$ by its corresponding vector of log-returns $\mathbf{x}_i, \mathbf{y}_i \in \mathcal{S}(\mathbb{R})$. We then evaluate the test statistic (53) where we choose $k : \mathbb{R}^d \times \mathbb{R}^d \rightarrow [0, +\infty)$ to be the Gaussian kernel (48) with $\sigma = 0.1$. We then compare the associated distribution of the MMD between the two histograms generated from the moment- and WK-means methods by reporting the similarity score from Definition 1.9.

Within-cluster evaluation is performed much in the same way as the between-cluster case: for either algorithm, and for each cluster \mathcal{C}_l , $l = 1, 2$ we draw $n \in \mathbb{N}$ pairwise samples $(\mu_i^1, \mu_i^2) \in \mathcal{C}_l \times \mathcal{C}_l$ and evaluate the biased MMD (53). We report the similarity score associated to the empirical distribution of each within-cluster MMD and plot the resulting histograms.

3.2.4. Cluster validation via the marginal MMD. Recall from Section 1 that the success of a clustering algorithm can be determined by appropriately balancing the trade-off between the *self-similarity* and *distinctness* of derived clusters in a given run. We begin with scores for each clustering algorithm via the indexes introduced in Definitions 3.3, 3.4, and 3.5. We provide the average silhouette coefficient \bar{S}_α with $\alpha = 0.2$. The results are presented in Table 1.

As noted in Section 3.2.3, scores associated to the first two indexes are not invariant under the choice of $(V, \|\cdot\|_V)$ and thus do not represent a like-for-like comparison. Yet we note

that the average Silhouette coefficient \bar{S}_α remains higher for the MK-means method than the WK-means, implying that (under the more traditional method of cluster validation) regimes clustered via the former belong to more appropriate clusters than the latter.

As outlined in Section 3.2.3, we applied our within- and between-cluster validation via the marginal MMD from Definition 1.7 by sampling $n = 1000000$ times from each cluster $\mathcal{C}_1, \mathcal{C}_2$ obtained from either method, and calculating the biased MMD (10). We order samples from clusters in ascending order to ensure like-for-like comparison between sample elements. Figure 3 shows two empirical distributions of the biased MMD between elements in the two clusters formed from the WK- and MK-means method.

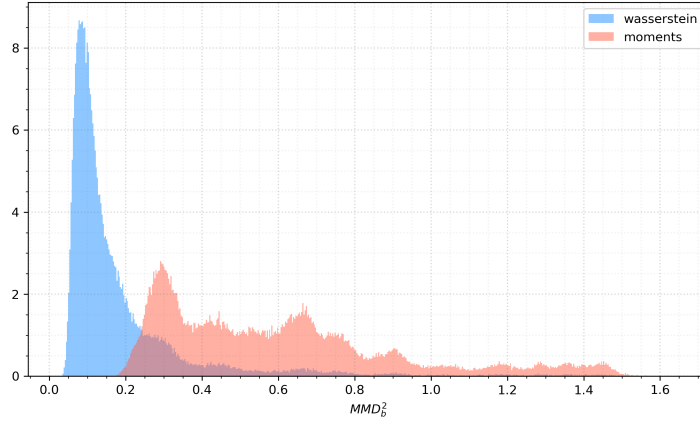


Figure 3. Histograms of between-cluster MMD approximation, Wasserstein vs moments method.

Figure 3, shows that the clusters obtained via the WK-means method are significantly more similar to each other than those obtained from the MK-means method. However, one cannot then conclude that the latter method provides a better clustering result if the disparity within groups is due to one cluster being composed primarily of outlier elements. Figure 4 gives the empirical distribution of the within-cluster MMD for each algorithm. These histograms were derived by calculating the biased MMD test statistic via the sub-sampling technique from Definition 1.9, with $n = 100000$.

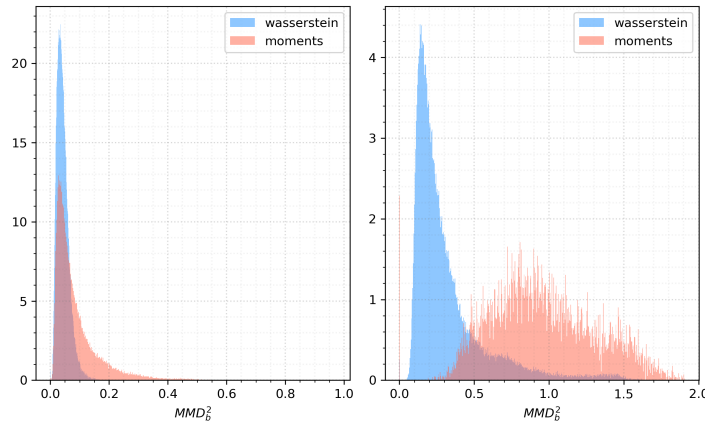


Figure 4. Histograms of within-cluster MMD approximation, Wasserstein vs moments method.

Algorithm	\mathcal{C}_1	\mathcal{C}_2
Moment	0.0631	1.1961
Wasserstein	0.0395	0.2304

Table 2. Self-similarity scores, WK- and MK-means algorithms.

Both Figure 4 and the self-similarity scores in Table 2 show that the clusters obtained via the WK-means algorithm are significantly more self-similar than those obtained from the MK-means algorithm.

3.3. Validation on synthetic data. Evaluating a given clustering algorithm on real market data is difficult for multiple reasons. One is that it is not possible to infer the underlying probabilistic structure associated to the stream of log-returns r^S that a clustering algorithm is run over, and thus one cannot say with any certainty what constitutes a “correct” clustering. A corollary to this is that it is impossible to know exactly at what point a regime change occurs when studying real market data.

Therefore, we evaluated both clustering algorithms on synthetic market data, where we specify beforehand at what times regime changes occur. Because we knew both the underlying probabilistic structure and the regime change periods *a priori*, we could further evaluate both how accurately either algorithm is classifying sequences of returns into regimes, and how closely the centroids $\{\bar{\mu}_l\}_{l=1,2}$ of each cluster correspond to the true distributions $\{\mathbb{P}_l\}_{l=1,2}$ associated to the synthetic data.

The methodology is as follows. For a given time interval $[0, T]$ with $T \in \mathbb{N}$, we define a mesh so that each time increment roughly represents one market hour. That is, with $n := 252 \times 7$, we set

$$\Delta = \left\{ \left[\frac{i-1}{n}, \frac{i}{n} \right] : i = 1, 2, \dots, nT \right\}.$$

Next, we define the number of regime changes $r \in \mathbb{N}$ we wish to observe. We specify their starting points and lengths by $(\tau_i, l_i) \in \mathbb{N} \times \mathbb{N}$ for $i = 1, \dots, r$, with

$$0 \leq \tau_0 < \tau_r + l_r \leq nT,$$

and

$$\tau_i + l_i + 2 < \tau_{i+1}, \quad \text{for } i = 1, \dots, r-1.$$

Each l_i can be a constant or a random variable. We thus obtain the set of disjoint intervals

$$(30) \quad R = \{[\tau_i, \tau_i + l_i] : i = 1, \dots, r\}$$

and their associated complements $N = \Delta \setminus R$ which partition the interval Δ into two sets. Intervals in R will correspond to times where we observe a regime change in our synthetic data, which will start at τ_i and end at $\tau_i + l_i$ for $i = 1, \dots, r$.

Once we have run a classification algorithm over a synthetic price path, we consider three measures of accuracy: total accuracy, accuracy during the standard regime (regime-off) and accuracy during the regime change (regime-on). This is calculated in the following way: for $i = 1, \dots, N-1$, associate to each log-return r_i^S the empirical measures $M_i = \{\mu_{j(i)}, \dots, \mu_{j(i)+v}\}$ it was a member of. With our chosen hyperparameters and $k = 2$, one has that $v \in [1, 5]$ and $j \in \mathbb{N}$ is the first measure that r_i^S is a member of. Note that if the overlap hyperparameter $h_2 = 0$ then $v = 1$. We then calculate which cluster each $\mu \in M_i$ is associated to, which gives us our predicted labels $\bar{y}^i = \{\bar{k}_1, \dots, \bar{k}_v\}$. We then aggregate

these labels into the row vector

$$\bar{Y}^i = \left(\sum_{j=1}^v \chi_{\{x=l\}}(\bar{k}_j) \right)_{l=1}^k \quad \text{for } i = 1, \dots, N-1,$$

where $k = 2$ is the number of clusters. In what follows we assume the assignment $\bar{k} = 1$ corresponds to the standard regime and $\bar{k} = 2$ the regime change. We then have the following definitions.

Definition 3.7. With the notation above, for a given vector of log-returns $r^s \in \mathcal{S}(\mathbb{R})$ and cluster assignments $\mathcal{C} = \{\mathcal{C}_l\}_{l=1}^k$, the *regime-off accuracy score (ROFS)* is given by

$$(31) \quad \text{ROFS}(r^s, \mathcal{C}) = \frac{\sum_{r_i^s \in N} \bar{Y}_1^i}{\sum_{r_i^s \in N} \sum_{k=1,2} \bar{Y}_k^i}.$$

Similarly, the *regime-on accuracy score (RONS)* is given by

$$(32) \quad \text{RONS}(r^s, \mathcal{C}) = \frac{\sum_{r_i^s \in R} \bar{Y}_2^i}{\sum_{r_i^s \in R} \sum_{k=1,2} \bar{Y}_k^i}.$$

Finally, *total accuracy (TA)* is given by

$$(33) \quad \text{TA}(r^s, \mathcal{C}) = \frac{\sum_{r_i^s \in N} \bar{Y}_1^i + \sum_{r_i^s \in R} \bar{Y}_2^i}{\sum_{i=1}^{N-1} \sum_{k=1,2} \bar{Y}_k^i}.$$

3.3.1. Geometric Brownian motion. In this section, we discuss how we tested both clustering algorithms on synthetic stock data which was modelled as a geometric Brownian motion (gBm). Let $\mathcal{M}(\Theta)$ be a family of models indexed by a parameter set $\Theta \subset \mathbb{R}^d$. Initially, we chose $\mathcal{M}(\Theta) = \text{gBm}(\mu, \sigma)$. We then specified two parameter combinations $\theta_{\text{bull}} = (\mu_1, \sigma_1)$ and $\theta_{\text{bear}} = (\mu_2, \sigma_2)$, corresponding to two market regimes.

We then construct a geometric Brownian motion with associated parameters θ_{bear} over intervals $[\tau_i, \tau_i + l_i] \in R$ for $i = 1, \dots, r$, and with parameters θ_{bull} elsewhere. We then run both clustering algorithms on the synthetic data and are returned the clusters with associated centroids as output. Since

$$(34) \quad \ln S_t \sim \text{Normal}((\mu - \sigma^2/2)t, \sigma^2 t) \quad \text{for all } t \geq 0,$$

the true measures $\{\mathbb{P}_l\}_{l=1,2}$ are given by

$$(35) \quad \mathbb{P}_l = \text{Normal}((\mu_l - \sigma_l^2/2)dt, \sigma_l^2 dt) \quad \text{for } l = 1, 2,$$

where $dt = 1/n$ is the mesh size. Due to the Gaussianity of the distribution of the true log-returns, it suffices to check the mean and variance of the centroid measures μ_l to gauge how close they are to the true measures \mathbb{P}_l for $l = 1, 2$.

We begin by testing on gBm paths with

$$\begin{aligned} \theta_{\text{bull}} &= (0.02, 0.2), & \text{and} \\ \theta_{\text{bear}} &= (-0.02, 0.3). \end{aligned}$$

We simulate a path over $T = 20$ years with $r = 10$ regime changes, and randomly chose each τ_i for $i = 1, \dots, 10$ and fixed $l_i = 0.5 \times 252 \times 7$. This choice corresponds to regime

changes persisting for approximately half a year. Our mesh grid is thus given by

$$\Delta = \left\{ \left[\frac{i-1}{1764}, \frac{i}{1764} \right] : i = 1, 2, \dots, 252 \times 7 \times 20 \right\}.$$

When a regime change occurs, the gBm parameters shift from θ_{bull} to θ_{bear} . Figure 5 shows an example of such a gBm path, with the regime change periods highlighted in red. Figure 5b gives the log-returns associated to Figure 5a, again with the regime changes highlighted in red.

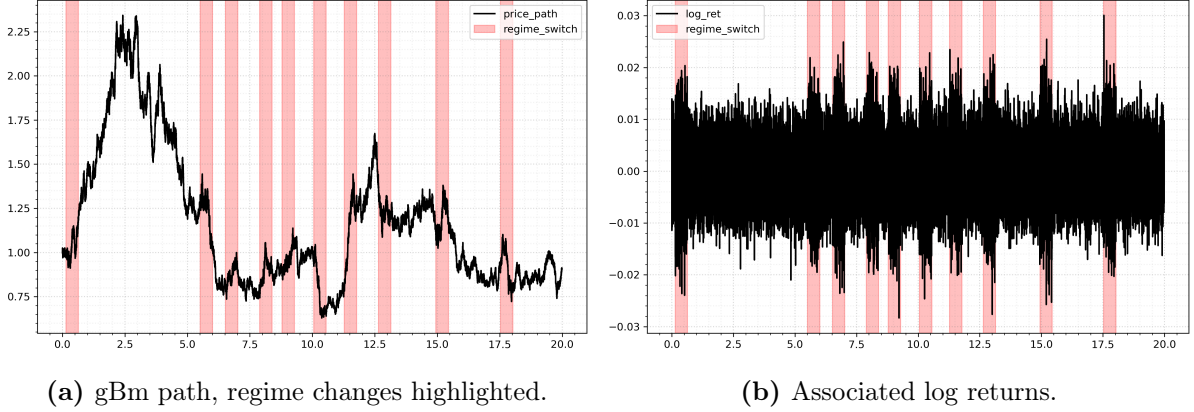


Figure 5. A gBm path and associated log returns.

We ran all three clustering algorithms on the same simulated path data. As seen from Figure 6 and Figure 7, both algorithms perform well - they are able to accurately capture both the regime changes and ends. We give a summary of the accuracy scores of each algorithm in Table 3 for a total of $n = 50$ runs.

Algorithm	Total	Regime-on	Regime-off	Runtime
Moment	93.23 \pm 0.41%	74.83 \pm 1.57%	99.38 \pm 0.1%	1.06 \pm 0.16s
HMM	58.16 \pm 7.11%	41.51 \pm 7.43%	63.72 \pm 11.94%	0.58 \pm 0.36s
Wasserstein	90.60 \pm 5.81%	87.24 \pm 4.11%	91.72 \pm 6.46%	0.87 \pm 0.16s

Table 3. Accuracy scores with 95% CI, gBm synthetic path, $n = 50$ runs.

It is interesting to note that, even in the Gaussian case, the WK-means algorithm does a better job of picking up regime changes than the standard approach. By comparison, the HMM tends to fail to detect the changes in regime at this fixed level of difference in parameter space and thus cannot determine regime change times. We provide the plots of the clustering algorithms in mean-variance space and the historical colouring plots in Figures 6 and 7.

We conclude this section by comparing the centroids obtained from either algorithm to the true measures. In this example, the distribution of the log-returns corresponding to either regime is given by

$$\begin{aligned} \mathbb{P}_{\theta_{\text{bull}}} &= \text{Normal}(-1.97 \times 10^{-21}, 2.27 \times 10^{-05}), \quad \text{and} \\ \mathbb{P}_{\theta_{\text{bear}}} &= \text{Normal}(-3.68 \times 10^{-05}, 5.1 \times 10^{-05}). \end{aligned}$$

Since the distribution of log-returns in this model are Gaussian, we study the mean and variance of our obtained centroids and compare these to the true values.

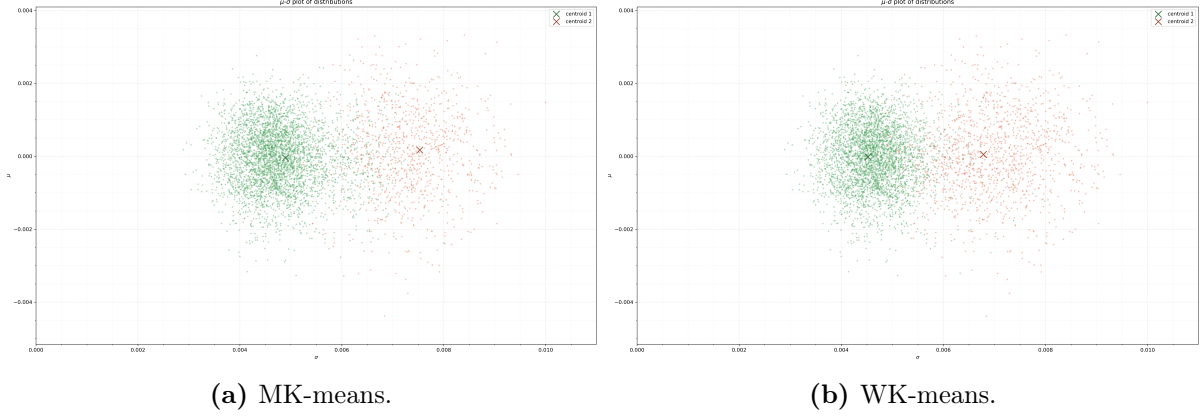


Figure 6. Plots of clusters in mean-variance space, synthetic gBm, example run.

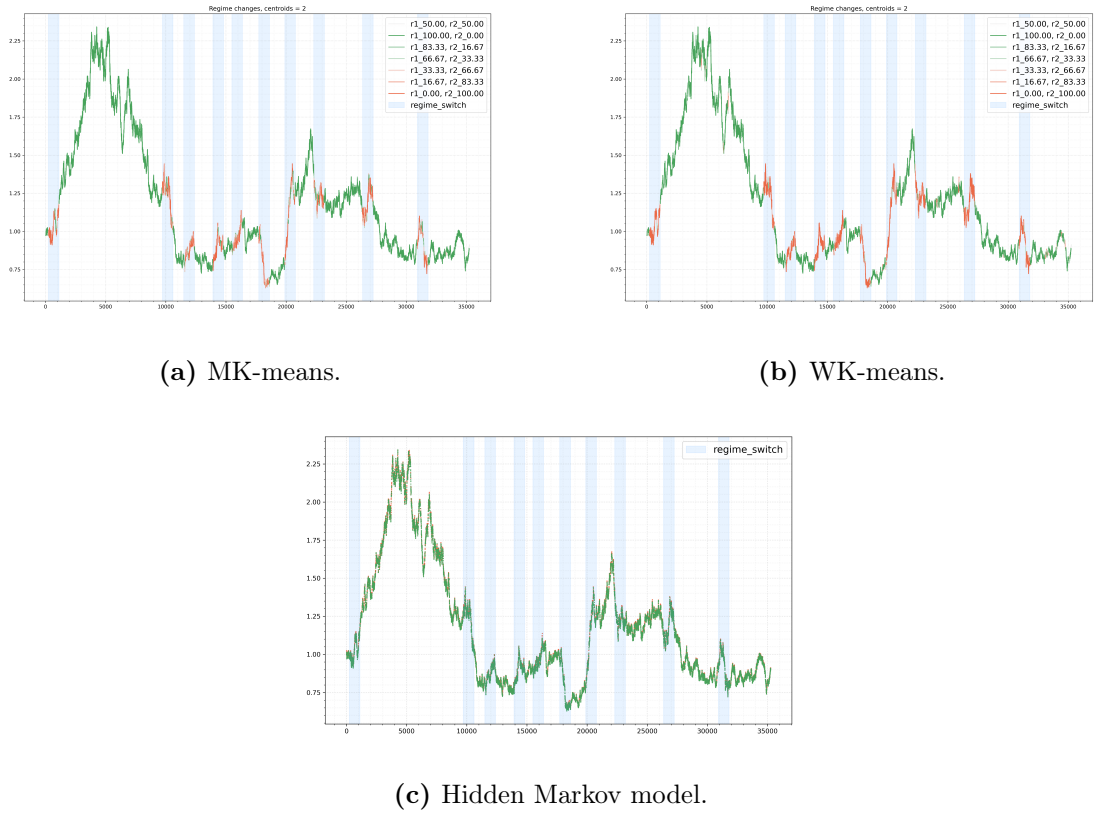


Figure 7. Historical cluster colouring plot, synthetic gBm, example run.

Figure 8 summarizes our findings. In the first row, we display the histogram of the partition means $\{\mathbb{E}[\mu_i]\}_{1 \leq i \leq M}$ along with solid lines representing the true means $\mathbb{E}[\mathbb{P}_{\theta_{\text{bull}}}]$ and $\mathbb{E}[\mathbb{P}_{\theta_{\text{bear}}}]$. In Figures 8a and 8b, the dashed lines represent the bull and bear centroid means corresponding to the MK-means and WK-means algorithms respectively. In the second row, we repeat the same procedure with the variances $\{\text{Var}(\mu_i)\}_{1 \leq i \leq M}$, their true values, and in Figures 8c and 8d the centroid variances corresponding to the MK-means and WK-means algorithms respectively.

We see that the centroids derived from either algorithm perform well as estimators for the true centroids. We note that it is not altogether unsurprising that the Wasserstein algorithm does not significantly outperform the moments-based method here: since we are

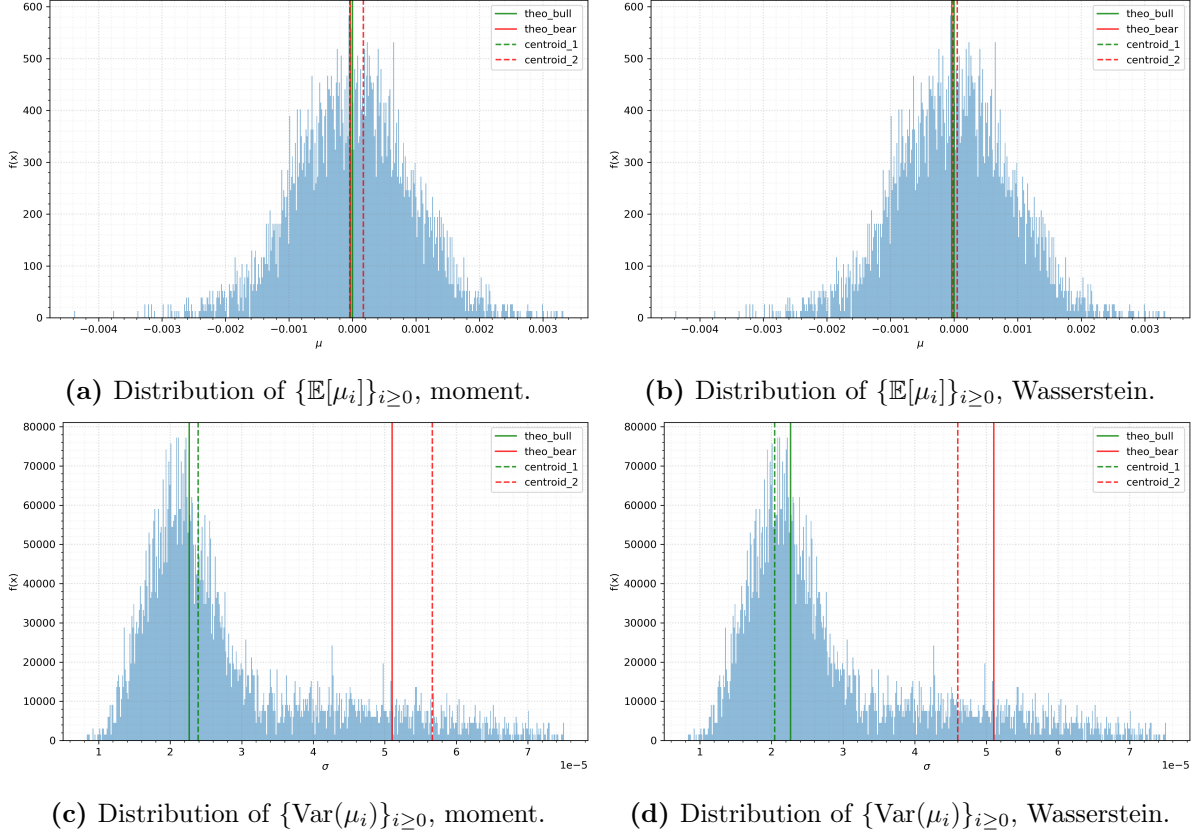


Figure 8. Approximations of mean and variance of true measures $\mathbb{P}_i, i = 1, 2$ by centroids of moments- and WK-means algorithms, gBm run.

working under the assumption that distributions of log-returns under the “true” model are completely determined by their mean and variance.

3.3.2. Merton jump diffusion. In this section, we outline a separate model used to generate synthetic data where the associated log-returns are non-Gaussian. In particular, we model stock prices by a Merton jump diffusion (MJD), which is given by the solution to the stochastic differential equation

$$(36) \quad dS_t = \mu S_t dt + \sigma S_t dW_t + S_{t-} dJ_t \quad \text{for } t \geq 0,$$

where

$$J_t = \sum_{j=1}^{N_t} V_j - 1.$$

Here, $N_t \sim \text{Po}(\lambda t)$ is a Poisson random variable, and $\ln(1 + V_j) \sim \text{Normal}(\gamma, \delta^2)$. Our model space $\mathcal{M}(\Theta)$ is given by

$$(37) \quad \mathcal{M}(\Theta) = \text{MJD}(\mu, \sigma, \lambda, \gamma, \delta) \quad \text{for } \Theta \subset \mathbb{R}^5.$$

The solution to eq. (36) is given by

$$(38) \quad S_t = F(0, t) \mathcal{E}(\sigma W_t) \prod_{j=1}^{N_t} V_j,$$

where $F(0, t) = S_0 \exp(\mu t)$, and $\mathcal{E} : \mathbb{R} \rightarrow [0, +\infty)$ is the Doléans-Dade stochastic exponential. Let $R_t^M = \ln(S_{t+dt}) - \ln(S_t)$ be the log-return associated to a realisation of (38)

at a time $t \in [0, T]$ on a mesh with grid size dt . Then, Synowiec (2008) gives that

$$(39) \quad \mathbb{E}[R_t^M] = ((\mu - \sigma^2/2) + \lambda\gamma)dt, \quad \text{and}$$

$$(40) \quad \text{Var}(R_t^M) = (\sigma^2 + \lambda(\delta^2 + \gamma^2))dt,$$

and we will use these quantities to check the suitability of the centroids from either algorithm to the true measures.

To test either algorithm on synthetic data as generated from eq. (37), we apply the same methodology as outlined in Subsection 3.3. That is, we define two sets of parameters $\theta_{\text{bear}}, \theta_{\text{bull}}$ and a partition Δ with regime changes $[\tau_i, \tau_i + l_i] \in R$ for $i = 1, \dots, r$, where R is given by (30). We then run both clustering algorithms over a MJD with parameters θ_{bear} over intervals in R and parameters θ_{bull} elsewhere. In regime switch dynamics are given by the parameter choices

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

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

We again set $r = 10$ and $l_i = 252 \times 7 \times 0.5$ for $i = 1, \dots, 10$. When a regime change occurs we shift the parameters of the MJD from θ_{bull} to θ_{bear} , and revert them back when the regime change ends. Figure 9 gives an example path and the associated log-returns, with the periods of regime change highlighted in red.

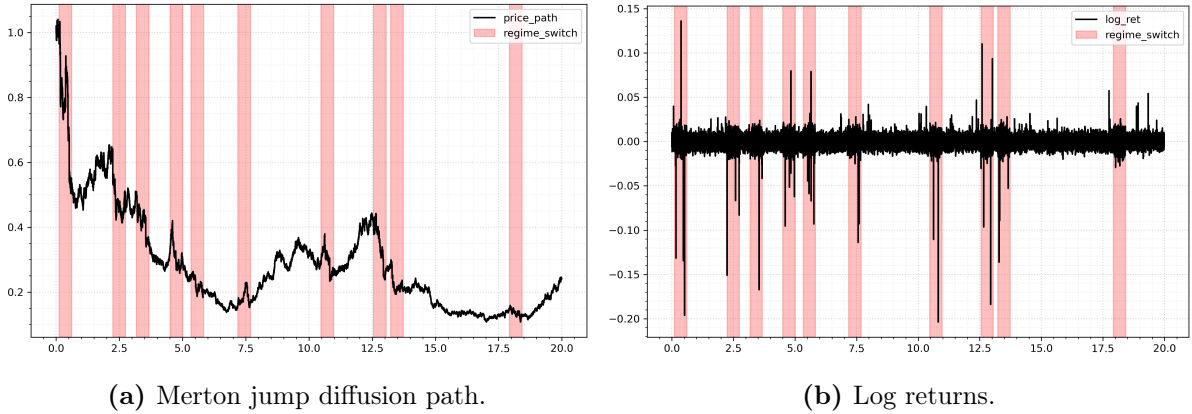


Figure 9. An iteration of a Merton jump diffusion path $\omega \rightarrow S_t^J$ and the associated log-returns, regime changes highlighted.

For the example path presented in Figure 9a, we present plots from all three algorithms where applicable. Figure 10 gives the projection of the derived clusters for the moment- and WK-means algorithm. Here, one can see that the MK-means approach fails to discern between the two market regimes as it is not robust enough to adjust for outlier return series in the bear regime. By comparison, the Wasserstein approach is relatively robust to these outliers, and is able to correctly identify the two different regimes.

Figure 11 shows the WK-means clusters in skew-kurtosis space. We see here that the algorithm correctly identifies several distributions exhibiting positive skew as belonging to the bull regime. The algorithm is also able to detect that distributions belonging to the bear regime are significantly more positively skewed.

We summarize the results for $n = 50$ runs with the parameters $(\theta_{\text{bull}}, \theta_{\text{bear}})$. It is clear that both the MK-means and HMM approach are unable to discern between periods of regime change and normalcy. Note that the high accuracy in the regime-on case for the HMM is due to the fact that it places all points into the first category. For the path in

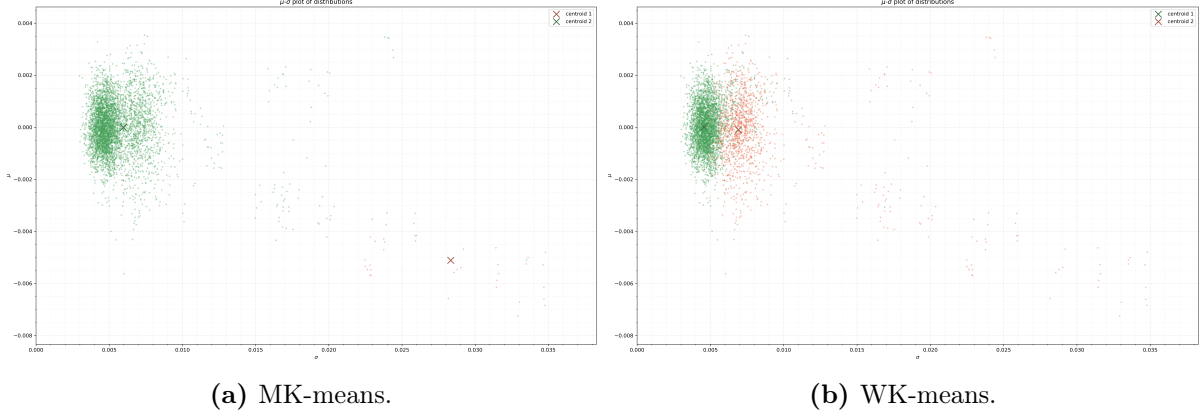


Figure 10. Outputs of clustering algorithms in mean-variance space, Merton jump diffusion, example run.

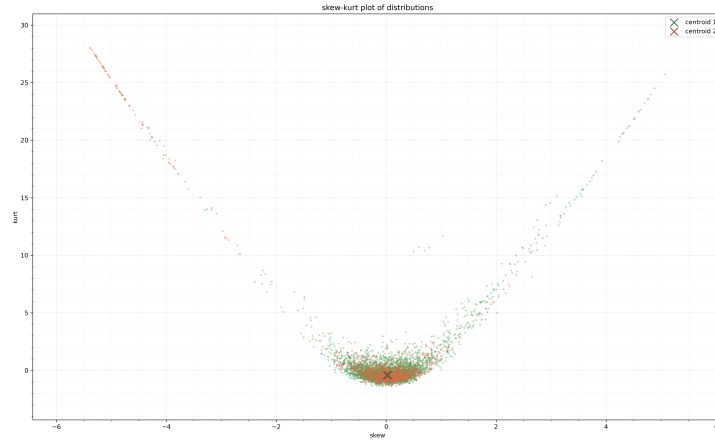


Figure 11. Clustered empirical Merton distributions in skew-kurtosis space, WK-means, example run.

Figure 9a, we give the historical colouring plot associated to a run of all three algorithms in Figure 12, which highlights the numerical results obtained from the table.

Algorithm	Total	Regime-on	Regime-off	Runtime
Moment	$66.64 \pm 3.42\%$	$27.25 \pm 8.73\%$	$79.79 \pm 7.40\%$	$1.71 \pm 0.28s$
HMM	$75.05 \pm 0.01\%$	$0.66 \pm 0.04\%$	$99.87 \pm 0.01\%$	$0.66 \pm 0.04s$
Wasserstein	$91.28 \pm 4.08\%$	$86.87 \pm 3.1\%$	$92.76 \pm 4.43\%$	$1.11 \pm 0.25s$

Table 4. Accuracy scores with 95% CI, Merton synthetic path, $n = 50$ runs.

As we did with the gBm example, we conclude the results section with a comparison between the mean and variance of the centroids obtained from either algorithm, and those associated to the true distributions as given by equations (39) and (40). As expected, from Figures 13a and 13c it is clear that the MK-means centroids do a poor job of approximating the true measures associated to the bull and bear regimes. We compare this to Figures 13b and 13d, where the mean and variance of the WK-means centroids are much closer to the theoretical counterparts.

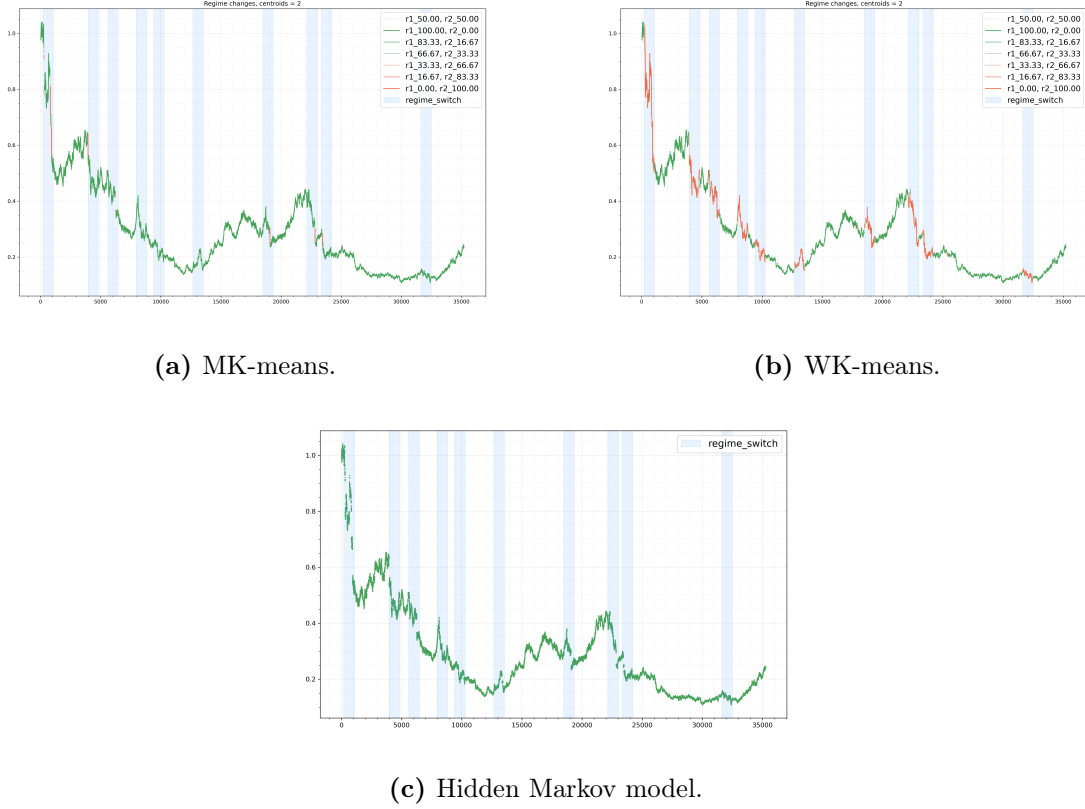


Figure 12. Historical cluster colouring on synthetic Merton jump diffusion price path, example runs.

3.4. Selection of hyperparameters. In this section we give a brief discussion regarding how the choice of hyperparameters (h_1, h_2) affects the results of the WK-means algorithm. We begin with a discussion to the first hyperparameter h_1 , which corresponds to the number of returns that form each empirical distribution within the clustering algorithm. Classically one would like to take as large a value of h_1 as is feasibly possible in order to best approximate empirically the true data-generating measure. In the regime clustering context, however, this is not always ideal: choosing h_1 to be too large can mean that regime changes are not captured, or (in a live data setting) the detection of such changes are lagged. However, certainly if h_1 is chosen too small spurious classifications dominated by noise will be made. Thus we believe that the choice of window length hyperparameter is more an art than a science and strongly depends on the application in mind.

Regarding the overlap hyperparameter h_2 : heuristically, a larger value of overlap parameter (relative to h_1) can be thought of in two ways. Mechanically it is a way of increasing the number of samples used in the clustering algorithm, which may be necessary in a low-data environment. Heuristically, by increasing the clustering set with measures that are very similar to each other, one is indirectly making a statement about how representative the observed sequence of log-returns (and thus clustering set measures) relative to what one might deem “standard” conditions. This phenomena can be seen in the simple case when one clusters on S&P 500 data before and after the GFC: for $k = 2$ and with $h_2 = 0$, one would expect that the outlier centroid $\bar{\mu}^2$ moves significantly faster to its new position than $\bar{\mu}^1$, whereas if h_2 is closer to h_1 , one expects the centroids to not initially

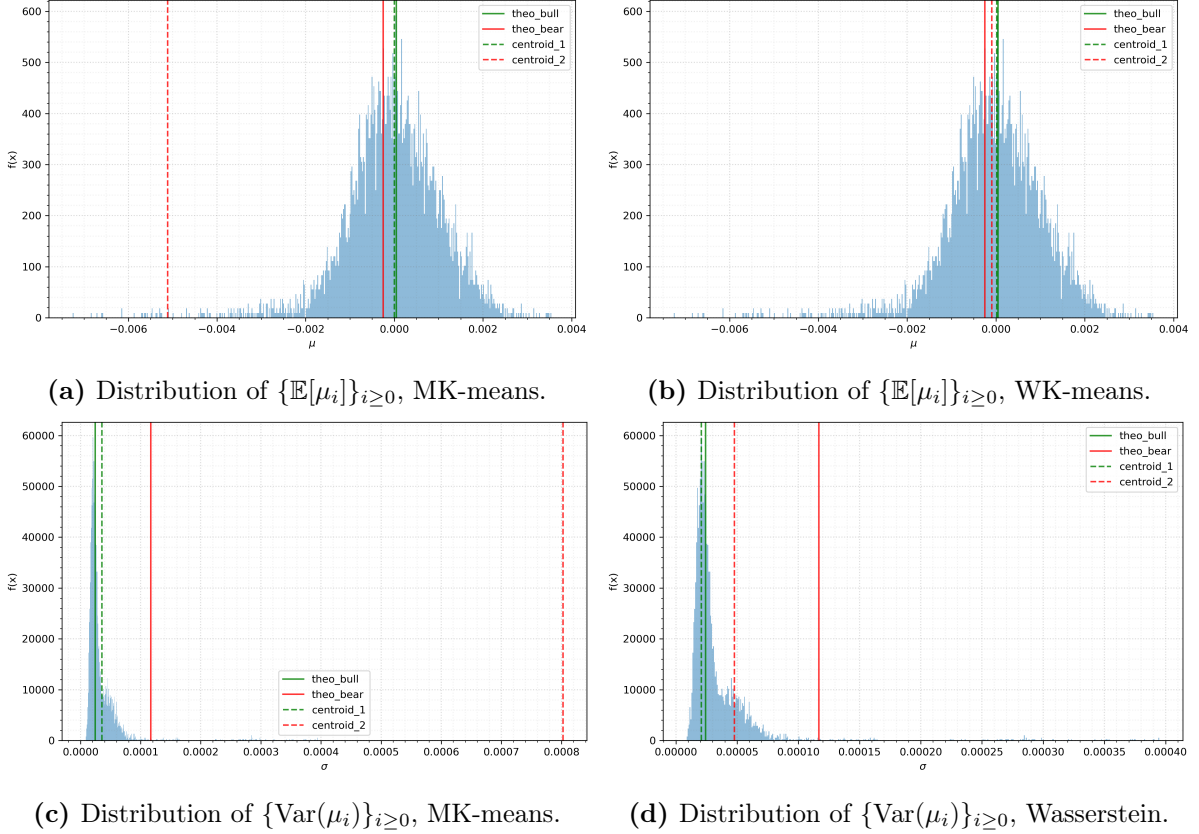


Figure 13. Approximations of mean and variance of true measures by centroids of moments- and WK-means algorithms.

change as much during the onset of the GFC, since new observations are less constituent relative to the corpus of measures preceding them.

We note however that in general the overlap hyperparameter does not have too large an effect on centroids obtained (and, thus, clusters) assuming that one is not operating in too low a data environment, and h_1 is suitably chosen. We present the results of clustering on SPY for the hyperparameter choice $h^1 = (35, 28)$, the choice we made in Subsection 3.2, and $h^2 = (35, 0)$ in Figure 14. Here, one can see that the obtained centroids from either algorithm do not change drastically in spite of the lower data density.

Indeed a simple KS two-sample test between the first centroids $(\bar{\mu}_1^{h_1}, \bar{\mu}_1^{h_2})$ returns a test statistic score of 0.02857 with an associated p -value of 1.0, and the second set of centroids $(\bar{\mu}_2^{h_1}, \bar{\mu}_2^{h_2})$ yields the same score and p -value.

Finally, we test the effect of changing the window length parameter h_1 . As stated in the beginning of the section, there exist many incorrect choices for h_1 in practice (too small, or too large), but reasonable choices will tend to give robust results. To show this, we chose a sequence of window lengths $H_1 = (7 + 7i)_{i=1}^10$ and set the overlap parameter $h_2 = \lfloor 3h_1/4 \rfloor$. We then ran the regime-switching experiment from Subsection 3.3 with the same switching dynamics as in Subsection 3.3.2, repeating the experiment 30 times for each hyperparameter choice. In particular, regime changes in this setting lasted for half a year, which amounts to 882 time steps given our year mesh, which means that only an unreasonably large rolling window would miss the given regime change. Figure 15 gives the various accuracy results (total, regime-on, and regime-off) for the window

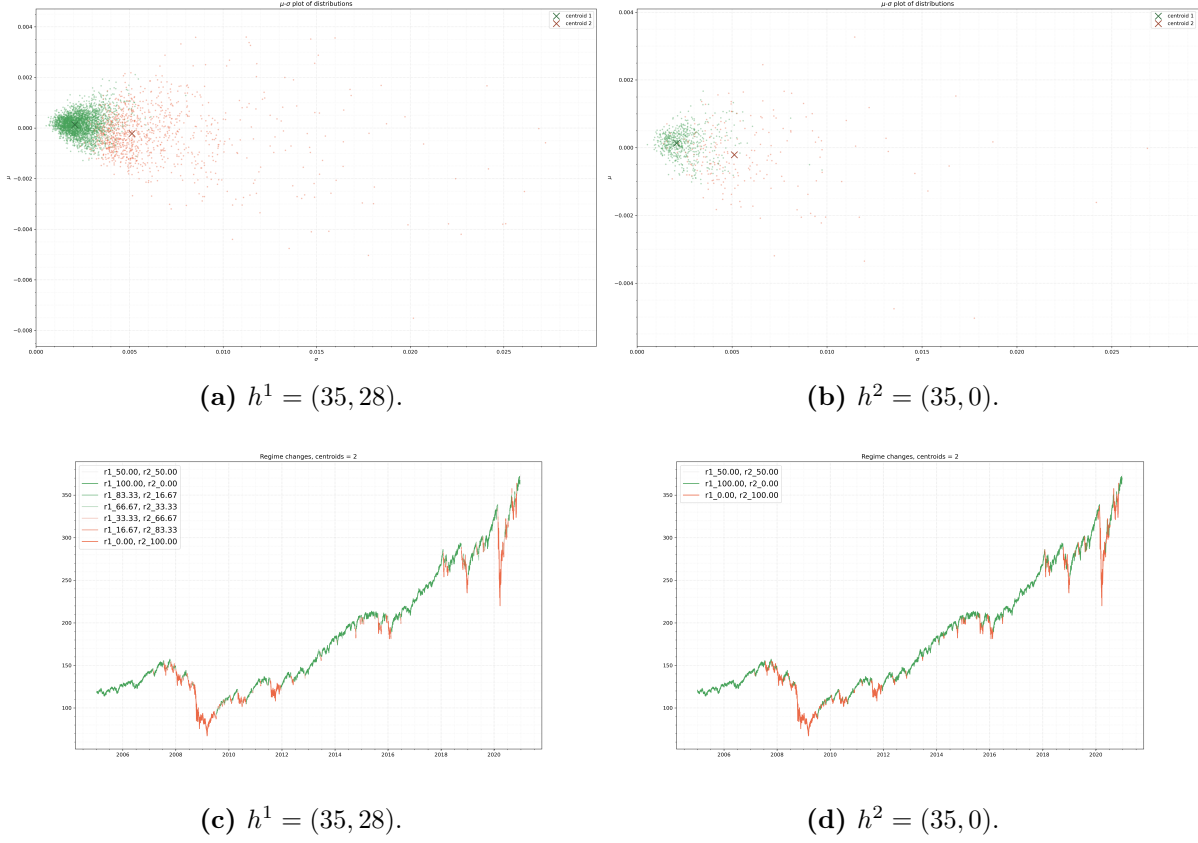


Figure 14. WK-means results with h^1 and h^2 hyperparameter choices, SPY price path.

length parameters given by H_1 . Past a certain threshold, the window length size does not matter and the model's accuracy converges.

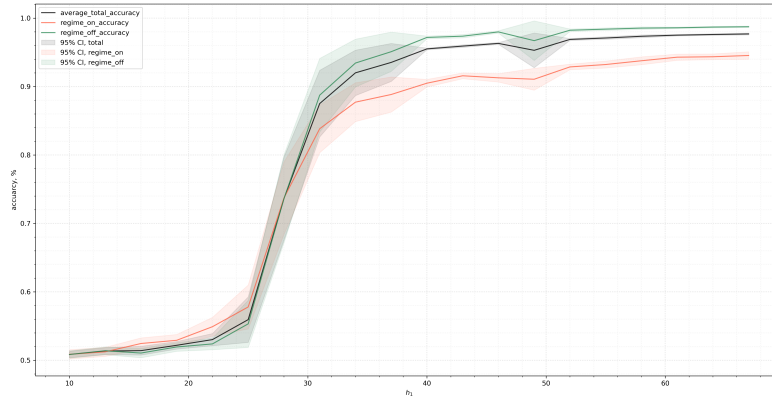


Figure 15. The effect of increasing h_1 on accuracy scores.

4. CONCLUSION

In this paper, we have shown that a slight modification to the k -means algorithm does an excellent job at classifying partitions of market returns into regimes. We have verified this on both real data by comparing to known periods of market instability, and on synthetic data where we explicitly determine when the regime changes occur. We have compared

this approach to a more standard moments-based algorithm which did not perform as well when returns were non-Gaussian, and a more classical approach using a hidden Markov model, which failed to accurately discern between regimes in the synthetic case. We also showed that clusters obtained via the Wasserstein approach are more self-similar than those derived from the moment-based method.

Future research would include employing other clustering algorithms rather than the standard k-means approach; for instance, fuzzy and hierarchical clustering methods. Further study into the robustness of the derived clusters under the choice of hyperparameters (that is, partitioning of the underlying time series) would also be relevant to understand how stable derived clusters are. We also note that there exist methodologies for determining the optimal number of clusters k to be used. Finally, a more analytic and rigorous study of the weak convergence of derived centroids to the true measures (in the synthetic data case) would also be of interest.

DECLARATIONS OF INTEREST

ZI was supported by EPSRC grant EP/R513064/1.

REFERENCES

- Achelis, S. B. (2001). *Technical analysis from a to z*. McGraw Hill New York.
- Ackermann, M., Blömer, J., & Sohler, C. (2008, 01). Clustering for metric and non-metric distance measures. In (Vol. 6, p. 799-808). doi: 10.1145/1347082.1347170
- Alquier, P., Chérif-Abdellatif, B.-E., Derumigny, A., & Fermanian, J.-D. (2020). *Estimation of copulas via maximum mean discrepancy*.
- Ambrosio, L., & Gigli, N. (2013). A user’s guide to optimal transport. In *Modelling and optimisation of flows on networks* (pp. 1–155). Springer.
- Ambrosio, L., Gigli, N., & Savare, G. (2005). *Gradient flows: In metric spaces and in the space of probability measures*. Birkhäuser Basel. Retrieved from <https://books.google.co.uk/books?id=HZqhWIq1-jgC>
- Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3), 337–404.
- Bayraktar, E., & Guo, G. (2021). Strong equivalence between metrics of wasserstein type. *Electronic Communications in Probability*, 26, 1–13.
- Berlinet, A., & Thomas-Agnan, C. (2011). *Reproducing kernel hilbert spaces in probability and statistics*. Springer US. Retrieved from <https://books.google.co.uk/books?id=bX3TBwAAQBAJ>
- Bonneel, N., Rabin, J., Peyré, G., & Pfister, H. (2015). Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1), 22–45.
- Bonnier, P., Kidger, P., Arribas, I. P., Salvi, C., & Lyons, T. J. (2019). Deep signatures. *CoRR*, abs/1905.08494. Retrieved from <http://arxiv.org/abs/1905.08494>
- Briol, F.-X., Barp, A., Duncan, A. B., & Girolami, M. (2019). *Statistical inference for generative models with maximum mean discrepancy*.
- Buehler, H., Horvath, B., Lyons, T., Perez Arribas, I., & Wood, B. (2020). A data-driven market simulator for small data environments. *Available at SSRN 3632431*.
- Cannon, R. L., Dave, J. V., & Bezdek, J. C. (1986). Efficient implementation of the fuzzy c-means clustering algorithms. *IEEE transactions on pattern analysis and machine intelligence*(2), 248–255.
- Chérif-Abdellatif, B.-E., & Alquier, P. (2021). *Finite sample properties of parametric mmd estimation: robustness to misspecification and dependence*.
- Cochrane, T., Foster, P., Lyons, T., & Arribas, I. P. (2020). Anomaly detection on streamed data. *arXiv preprint arXiv:2006.03487*.
- Cont, R. (2001). Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative finance*, 1(2), 223.
- Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2), 224-227. doi: 10.1109/TPAMI.1979.4766909
- Dias, J. G., Vermunt, J. K., & Ramos, S. (2015). Clustering financial time series: New insights from an extended hidden markov model. *European Journal of Operational Research*, 243(3), 852–864.
- Dunn, J. C. (1974). Well-separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, 4(1), 95-104. Retrieved from <https://doi.org/10.1080/01969727408546059> doi: 10.1080/01969727408546059
- Fukumizu, K., Gretton, A., Schölkopf, B., & Sriperumbudur, B. K. (2009). Characteristic kernels on groups and semigroups. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Advances in neural information processing systems* (Vol. 21). Curran Associates, Inc. Retrieved from <https://proceedings.neurips.cc/paper/>

- 2008/file/d07e70efcfab08731a97e7b91be644de-Paper.pdf
- Fukumizu, K., Gretton, A., Sun, X., & Schölkopf, B. (2007). Kernel measures of conditional dependence. In *Nips* (Vol. 20, pp. 489–496).
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. The MIT Press.
- Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., & Smola, A. J. (2007). A kernel method for the two-sample-problem. In B. Schölkopf, J. C. Platt, & T. Hoffman (Eds.), *Advances in neural information processing systems 19* (pp. 513–520). MIT Press. Retrieved from <http://papers.nips.cc/paper/3110-a-kernel-method-for-the-two-sample-problem.pdf>
- Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., & Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar), 723–773.
- Gretton, A., Fukumizu, K., Harchaoui, Z., & Sriperumbudur, B. K. (2009). A fast, consistent kernel two-sample test. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems 22* (pp. 673–681). Curran Associates, Inc. Retrieved from <http://papers.nips.cc/paper/3738-a-fast-consistent-kernel-two-sample-test.pdf>
- Guidolin, M. (2011). Markov switching models in empirical finance. In *Missing data methods: Time-series methods and applications*. Emerald Group Publishing Limited.
- Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica: Journal of the econometric society*, 357–384.
- Henderson, K., Gallagher, B., & Eliassi-Rad, T. (2015). Ep-means: An efficient nonparametric clustering of empirical probability distributions. In *Proceedings of the 30th annual acm symposium on applied computing* (pp. 893–900).
- Horvath, B., Muguruza, A., & Tomas, M. (2019). Deep learning volatility. Available at SSRN 3322085.
- Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A. Y. (2002). An efficient k-means clustering algorithm: Analysis and implementation. *IEEE transactions on pattern analysis and machine intelligence*, 24(7), 881–892.
- Kolouri, S., Nadjahi, K., Simsekli, U., Badeau, R., & Rohde, G. K. (2019). Generalized sliced wasserstein distances. *arXiv preprint arXiv:1902.00434*.
- Kondratyev, A., Schwarz, C., & Horvath, B. (2020). Data anonymisation, outlier detection and fighting overfitting with restricted boltzmann machines. *Outlier Detection and Fighting Overfitting with Restricted Boltzmann Machines (January 27, 2020)*.
- Kuan, C.-M. (2002). Lecture on the markov switching model. *Institute of Economics Academia Sinica*, 8(15), 1–30.
- Lahmiri, S. (2016). Clustering of casablanca stock market based on hurst exponent estimates. *Physica A: Statistical Mechanics and its Applications*, 456, 310–318.
- Lange, T., & Rahbek, A. (2009). An introduction to regime switching time series models. In *Handbook of financial time series* (pp. 871–887). Springer.
- Li, J., & Wang, J. Z. (2008). Real-time computerized annotation of pictures. *IEEE transactions on pattern analysis and machine intelligence*, 30(6), 985–1002.
- Lux, T., & Marchesi, M. (2000). Volatility clustering in financial markets: a microsimulation of interacting agents. *International journal of theoretical and applied finance*, 3(04), 675–702.
- Maheu, J., McCurdy, T., & Song, Y. (2012, 07). Components of bull and bear markets: Bull corrections and bear rallies. *Journal of Business and Economic Statistics*, 30. doi: 10.2139/ssrn.1939486

- Mason, D. M., & Schuenemeyer, J. H. (1983). A modified kolmogorov-smirnov test sensitive to tail alternatives. *The annals of Statistics*, 933–946.
- Mi, L., Zhang, W., Gu, X., & Wang, Y. (2018). Variational wasserstein clustering. In *Proceedings of the european conference on computer vision (eccv)* (pp. 322–337).
- Ni, H., Szpruch, L., Wiese, M., Liao, S., & Xiao, B. (2020). *Conditional sig-wasserstein gans for time series generation*.
- Nielsen, F., & Nock, R. (2009). Sided and symmetrized bregman centroids. *IEEE transactions on Information Theory*, 55(6), 2882–2904.
- Nielsen, F., Nock, R., & Amari, S.-i. (2014). On clustering histograms with k-means by using mixed α -divergences. *Entropy*, 16(6), 3273–3301.
- Niu, Y. S., Hao, N., & Zhang, H. (2016). Multiple change-point detection: a selective overview. *Statistical Science*, 611–623.
- Pelletier, D. (2006). Regime switching for dynamic correlations. *Journal of Econometrics*, 131(1), 445–473. Retrieved from <http://www.sciencedirect.com/science/article/pii/S0304407605000187> doi: <https://doi.org/10.1016/j.jeconom.2005.01.013>
- Rabin, J., Peyré, G., Delon, J., & Bernot, M. (2011). Wasserstein barycenter and its application to texture mixing. In *International conference on scale space and variational methods in computer vision* (pp. 435–446).
- Ray, S., & Turi, R. H. (1999). Determination of number of clusters in k-means clustering and application in colour image segmentation. In *Proceedings of the 4th international conference on advances in pattern recognition and digital techniques* (pp. 137–143).
- Revuz, D., & Yor, M. (2004). *Continuous martingales and brownian motion*. Springer Berlin Heidelberg. Retrieved from <https://books.google.co.uk/books?id=1m195FLM5koC>
- Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53–65. Retrieved from <https://www.sciencedirect.com/science/article/pii/0377042787901257> doi: [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7)
- Santambrogio, F. (2015). *Optimal transport for applied mathematicians. calculus of variations, pdes and modeling*.
- Steinwart, I. (2001). On the influence of the kernel on the consistency of support vector machines. *Journal of machine learning research*, 2(Nov), 67–93.
- Synowiec, D. (2008). Jump-diffusion models with constant parameters for financial log-return processes. *Computers and Mathematics with Applications*, 56(8), 2120–2127. Retrieved from <http://www.sciencedirect.com/science/article/pii/S0898122108003477> doi: <https://doi.org/10.1016/j.camwa.2008.02.051>
- Veldhuis, R. (2002). The centroid of the symmetrical kullback-leibler distance. *IEEE signal processing letters*, 9(3), 96–99.
- Wang, J., Gao, R., & Xie, Y. (2021). *Two-sample test using projected wasserstein distance: Breaking the curse of dimensionality*.
- Ye, J., Wu, P., Wang, J. Z., & Li, J. (2017). Fast discrete distribution clustering using wasserstein barycenter with sparse support. *IEEE Transactions on Signal Processing*, 65(9), 2317–2332.

APPENDIX

APPENDIX A. THE k -MEANS ALGORITHM

In this section, we outline some of the notation used in the paper, along with some standard results regarding the classical k -means algorithm.

Recall that $X \in \mathcal{S}(V)$ is a stream of data over a normed vector space $(V, \|\cdot\|_V)$.

Definition A.1 (Set of clusterings over X). We write

$$\mathcal{C}(X) = \left\{ \{\mathcal{C}_i\}_{0 \leq i \leq n} : \mathcal{C}_i \cap \mathcal{C}_j = \emptyset, \bigcup_{i=1}^n \mathcal{C}_i = X, n \in \mathbb{N} \right\}$$

to be the set of all possible (disjoint) clusterings over X .

The k -means algorithm returns an element $\mathcal{C}^* \in \mathcal{C}(X)$ which is locally optimal with respect to the induced metric $d : V \times V \rightarrow [0, +\infty)$ on V .

Before continuing with a more detailed explanation of the k -means algorithm, we introduce the following definitions.

Definition A.2 (Within-cluster variation). Let $k \in \mathbb{N}$ and let $X \in \mathcal{S}(V)$ be a stream of data over a normed vector space V . Suppose $\mathcal{C} \subset \mathcal{C}(X)$ are disjoint clusters over X . Associate to each \mathcal{C}_l its centroid \bar{x}_l for $l = 1, \dots, k$. Then, for a given \mathcal{C}_l , the *within-cluster variation* is defined as

$$(41) \quad \text{WC}(\mathcal{C}_l) = \sum_{x \in \mathcal{C}_l} \|x - \bar{x}_l\|_V^2 \quad \text{for } l = 1, \dots, k.$$

Definition A.3 (Total-cluster variation). With the notation of Definition A.2, define

$$(42) \quad \text{TC}(\mathcal{C}) = \sum_{i=1}^k \text{WC}(\mathcal{C}_i)$$

to be the *total-cluster variation* corresponding to a clustering $\mathcal{C} \in \mathcal{C}(X)$ on the normed vector space $(V, \|\cdot\|_V)$.

Recall that $\{\mathcal{C}_l^n\} \in \mathcal{C}(X)$ denotes an intermediate disjoint clustering of the set X at step $n \in \mathbb{N}$. The k -means algorithm first assigns nearest neighbours via (7) with respect to V . The next step in the algorithm is to update the centroids via an aggregation function $\alpha : 2^V \rightarrow V$ which gives a central element $\bar{x}_l \in V$ from the set of nearest neighbours \mathcal{C}_l for $l = 1, \dots, k$. In the classical k -means on \mathbb{R}^d , this function is given by

$$(43) \quad \alpha(\mathcal{C}_l) = \left(\frac{1}{|\mathcal{C}_l|} \sum_{x \in \mathcal{C}_l} x_j \right)_{1 \leq j \leq d}.$$

Here, $|\mathcal{C}|$ denotes the cardinality of the set \mathcal{C} . Note that other choices of $(V, \|\cdot\|_V)$ necessitate different aggregation methods depending on the structure of V .

Continuing, the algorithm then updates the centroids via the function α :

$$\bar{x}_l^n = \alpha(\mathcal{C}_l^n) \quad \text{for } l = 1, \dots, k.$$

The new centroids \bar{x}^n are then compared to \bar{x}^{n-1} via the following stopping rule.

Definition A.4 (*k*-means stopping rule). Suppose $(V, \|\cdot\|_V)$ is a normed vector space. For fixed $k \in \mathbb{N}$, consider a loss function $l : V^k \times V^k \rightarrow \mathbb{R}_+$ given by

$$(44) \quad l(x, y) = \sum_{i=1}^k \|x_i - y_i\|_V,$$

For a tolerance level $\varepsilon > 0$, the *stopping rule* corresponding to the standard *k*-means algorithm is given by

$$(45) \quad l(\bar{x}^{n-1}, \bar{x}^n) < \varepsilon,$$

where $n \in \mathbb{N}$ denotes the step of the algorithm, and we take $V = \mathbb{R}^d$.

Remark A.5. Various algorithms which attempt to find the optimal number of clusters k that should be used to separate data X consider (41) as the loss to be minimised over all possible clusterings. We will not cover these algorithms in this paper (see, for instance (Ray & Turi, 1999)), and their applications to the MRCP are topics for future research.

A given run of the *k*-means algorithm is characterised by the 2-tuple of centroids and nearest neighbour assignments $(\bar{x}_l, \mathcal{C}_l)_{1 \leq l \leq k}$. We conclude this section with the following proposition.

Proposition A.6. *The k-means algorithm converges in finitely many steps to a local minima.*

Proof. Finiteness is guaranteed since the number of possible partitions of datum X is at most k^N . Thus, the function $\text{TC} : \mathcal{C}(X) \rightarrow [0, +\infty]$ necessarily achieves a global minimum. Therefore, the sequence $(\text{TC}(\mathcal{C}^n))_{n \geq 1}$ is non-increasing (by definition of the *k*-means update step (7)) and bounded from below, which guarantees convergence to a local minima. \square

Finally, we summarise Section 1.4 and Appendix A with Algorithm 1.

Algorithm 1: Standard *k*-means algorithm

Result: k centroids
initialise centroids by sampling k times from X ;
while *loss_function* > *tolerance* **do**
 foreach x_i **do**
 | **assign** closest centroid wrt Euclidean distance;
 end
 update centroids;
 calculate *loss_function*;
end

APPENDIX B. THE MAXIMUM MEAN DISCREPANCY

In this section, we include extra details regarding the derivation of the MMD from Definition 1.7. In particular we show how one can show the MMD can be employed as a metric on the space of probability measures.

Definition B.1 (Reproducing kernel Hilbert space, (Aronszajn, 1950), Section 1.1). Suppose \mathcal{X} is a non-empty set, and let $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ be a Hilbert space of functions $f : \mathcal{X} \rightarrow \mathbb{R}$. We call a positive definite function $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ a *reproducing kernel* of \mathcal{H} if

(i) For all $x \in \mathcal{X}$, we have that $k(\cdot, x) \in \mathcal{H}$, and

(ii) For all $x \in \mathcal{X}$ and $f \in \mathcal{H}$, one has that

$$(46) \quad f(x) = \langle f(\cdot), k(\cdot, x) \rangle_{\mathcal{H}},$$

referred to as the *reproducing property*.

We call the Hilbert space \mathcal{H} associated to k a *reproducing kernel Hilbert space* (RKHS).

We can associate to each RKHS \mathcal{H} the *canonical feature map* given by $\phi(x) = k(\cdot, x)$. We thus have that

$$k(x, y) = \langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{H}} = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}} \quad \text{for all } x, y \in \mathcal{X},$$

by the reproducing property from Definition B.1. Directly from the definition of a RKHS \mathcal{H} , we have the following equivalent definition.

Definition B.2 ((Berlinet & Thomas-Agnan, 2011), Theorem 1). Suppose \mathcal{H} is a Hilbert space. Define $\delta_x : \mathcal{H} \rightarrow \mathbb{R}$ to be the evaluation map. Then, \mathcal{H} is a RKHS if and only if δ_x is continuous.

Proof. Suppose that \mathcal{H} is a RKHS. Denote by \mathcal{H}' as the dual of \mathcal{H} . One has that

$$(47) \quad \begin{aligned} |\delta_x(f)| &= |f(x)| = |\langle f, k(\cdot, x) \rangle_{\mathcal{H}}| \\ &\leq \|f\|_{\mathcal{H}} \langle k(\cdot, x), k(\cdot, x) \rangle_{\mathcal{H}}^{1/2} \\ &= \sqrt{k(x, x)} \|f\|_{\mathcal{H}}, \end{aligned}$$

so in particular the linear operator δ_x is bounded with operator norm equal to $\sqrt{k(x, x)}$, which is well-defined by positive definiteness of k . Since the upper bound in (47) is achieved by $f = k(\cdot, x)$, δ_x is bounded with operator norm $\|\delta_x\|_{\mathcal{H}'} = \sqrt{k(x, x)}$. Since δ_x is bounded, it is continuous.

Now suppose that δ_x is bounded, so $\delta_x \in \mathcal{H}'$. By the Riesz representation theorem, there exists an element $f_{\delta_x} \in \mathcal{H}$ such that $\delta_x(f) = \langle f, f_{\delta_x} \rangle_{\mathcal{H}}$. Define $k(x, x') := f_{\delta_x}(x')$. We then have that $\delta_x(f) = f(x) = \langle f, k(\cdot, x) \rangle_{\mathcal{H}}$ and $k(\cdot, x) \in \mathcal{H}$ by construction. Thus, properties (1) and (2) are satisfied in Definition B.1, so \mathcal{H} is a RKHS. \square

In what follows, we will choose our function class \mathcal{F} from Definition 1.7 to be the unit ball in a RKHS \mathcal{H} with associated reproducing kernel

$$(48) \quad k(x, y) = \exp\left((2\sigma)^{-2} \|x - y\|_{\mathbb{R}^d}^2\right) \quad \text{for } \sigma > 0,$$

called the *Gaussian kernel*. Importantly, such a RKHS has the following property, as shown in Steinwart (Steinwart, 2001).

Definition B.3 (Steinwart (Steinwart, 2001), Definition 4). Let (\mathcal{X}, d) be a compact metric space. Suppose \mathcal{H} is a RKHS with associated reproducing kernel $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$. We call \mathcal{H} *universal* if

(i) $k(\cdot, \cdot)$ is continuous, and

(ii) \mathcal{H} is dense in $C_b(\mathcal{X})$, the space of bounded continuous functions on \mathcal{X} , with respect to the supremum norm $\|\cdot\|_{\infty}$.

This choice of RKHS ensures that the MMD is metric on the space of Borel probability measures, which allows us to conclude the following.

Theorem B.4 ((Gretton et al., 2012), Theorem 5). *Let \mathcal{F} be the unit ball of a universal RKHS \mathcal{H} comprised of \mathbb{R} -functions on a compact space \mathcal{X} . Suppose $\mu, \nu \in \mathcal{P}(\mathcal{X})$ are Borel. Then $\text{MMD}[\mathcal{F}, \mu, \nu] = 0$ if and only if $\mu = \nu$.*

Proof. For $\mu \in \mathcal{P}(\mathcal{X})$, consider the linear functional $T_\mu : \mathcal{F} \rightarrow \mathbb{R}$ given by $T_\mu(f) = \mathbb{E}_\mu[f]$. We have that

$$|T_\mu(f)| = |\mathbb{E}_\mu[f(x)]| \leq \mathbb{E}_\mu[|f(x)|] = \mathbb{E}_\mu[|\langle f, k(\cdot, x) \rangle_{\mathcal{H}}|] \leq \mathbb{E}_\mu[\sqrt{k(x, x)}] \|f\|_{\mathcal{H}},$$

so in particular T_μ is continuous if $k(\cdot, \cdot)$ is measurable and $\mathbb{E}_\mu[\sqrt{k(x, x)}] < +\infty$. By the Riesz representation theorem, there exists a $m_\mu \in \mathcal{H}$ such that $T_\mu(f) = \langle f, m_\mu \rangle_{\mathcal{H}}$. In particular,

$$m_\mu(x) = \langle m_\mu, k(\cdot, x) \rangle_{\mathcal{H}} = \mathbb{E}_\mu[k(\cdot, x)] = \mathbb{E}_\mu[\phi(x)].$$

We call m_μ the *mean embedding* of μ in \mathcal{H} . From (9), we have that

$$\begin{aligned} \text{MMD}^2[\mathcal{F}, \mu, \nu] &= \sup_{f \in \mathcal{F}} \left(\mathbb{E}_\mu[f(x)] - \mathbb{E}_\nu[f(y)] \right)^2 \\ &= \sup_{f \in \mathcal{F}} \left(\langle f, m_\mu \rangle_{\mathcal{H}} - \langle f, m_\nu \rangle_{\mathcal{H}} \right)^2 \\ (49) \quad &= \sup_{\|f\|_{\mathcal{H}} \leq 1} \left(\langle m_\mu - m_\nu, f \rangle_{\mathcal{H}} \right)^2 = \|m_\mu - m_\nu\|_{\mathcal{H}}^2. \end{aligned}$$

Here, we have used the fact that \mathcal{F} is a unit ball in \mathcal{H} . Suppose that $\mu = \nu$. By (49), this implies $\text{MMD}[\mathcal{F}, \mu, \nu] = 0$.

Now suppose that $\mu \neq \nu$. By universality of \mathcal{H} , for any $\varepsilon > 0$ and $f \in C_b(\mathcal{X})$ there exists a $g \in \mathcal{H}$ such that

$$(50) \quad \|f - g\|_{\infty} < \frac{\varepsilon}{2}.$$

Then, we have that

$$(51) \quad |\mathbb{E}_\mu[f] - \mathbb{E}_\nu[f]| \leq |\mathbb{E}_\mu[f] - \mathbb{E}_\mu[g]| + |\mathbb{E}_\mu[g] - \mathbb{E}_\nu[g]| + |\mathbb{E}_\nu[g] - \mathbb{E}_\nu[f]| < \frac{\varepsilon}{2} + 0 + \frac{\varepsilon}{2} = \varepsilon,$$

where we have used the fact that for measures μ and ν , (50) gives that

$$|\mathbb{E}[f] - \mathbb{E}[g]| \leq \mathbb{E}[|f(x) - g(x)|] \leq \|f - g\|_{\infty},$$

and

$$|\mathbb{E}_\mu[g] - \mathbb{E}_\nu[g]| = |\langle g, m_\mu - m_\nu \rangle_{\mathcal{H}}| = 0$$

since we assumed $\text{MMD}[\mathcal{F}, \mu, \nu] = 0$. Thus $\mu = \nu$ as (51) holds for all $f \in C_b(\mathcal{X})$. \square

Remark B.5. Suppose (X, Σ) is a general measurable space (and not necessarily compact). Recalling Definition (1.8), if a kernel k associated to the RKHS \mathcal{H} is *characteristic*, so the mapping

$$\mathcal{P}(\mathcal{X}) \ni \mu \mapsto \mathbb{E}_{X \sim \mu}[k(\cdot, X)] \in \mathbb{R}$$

is injective, then one can conclude Theorem B.4 via equation (49).

Using the definition of the mean embedding, the fact that $m_\mu(t) = \mathbb{E}_{x \sim \mu}[k(t, x)]$, and the reproducing property of \mathcal{F} , we can write (49) as

$$(52) \quad \text{MMD}^2[\mathcal{F}, \mu, \nu] = \mathbb{E}_{x, x' \sim \mu}[k(x, x')] - 2\mathbb{E}_{x \sim \mu, y \sim \nu}[k(x, y)] + \mathbb{E}_{y, y' \sim \nu}[k(y, y')]$$

since, for example,

$$\langle m_\mu, m_\mu \rangle_{\mathcal{H}} = \mathbb{E}_{x \sim \mu}[m_\mu(x)] = \mathbb{E}_{x, x' \sim \mu}[k(x, x')].$$

Given samples $x = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_m)$, a biased empirical estimate of (52) is given by

$$(53) \quad \text{MMD}_b[\mathcal{F}, x, y] = \left[\frac{1}{n^2} \sum_{i,j=1}^n k(x_i, x_j) - \frac{2}{mn} \sum_{i,j=1}^{m,n} k(x_i, y_j) + \frac{1}{m^2} \sum_{i,j=1}^m k(y_i, y_j) \right]^{\frac{1}{2}}.$$

We will use the test statistic (53) to evaluate the success of a given clustering algorithm.

APPENDIX C. THE WASSERSTEIN DISTANCE

In this section, we include proofs of results regarding the Wasserstein distance.

Proposition C.1 (Wasserstein barycenter, empirical measures). *Suppose that $\{\mu_i\}_{1 \leq i \leq M}$ are a family of empirical probability measures, each with N atoms $\alpha_i^1, \dots, \alpha_i^N$ for $i = 1, \dots, M$. Let*

$$a_j = \text{Median}(\alpha_1^j, \dots, \alpha_M^j) \quad \text{for } j = 1, \dots, N.$$

Then, the cumulative distribution function of the Wasserstein barycenter $\bar{\mu} \in \mathcal{P}_p(\mathbb{R})$ over $\{\mu_i\}_{1 \leq i \leq M}$ with respect to the 1-Wasserstein distance is given by

$$(54) \quad \bar{\mu}((-\infty, x]) = \frac{1}{N} \sum_{i=1}^N \chi_{a_i \leq x}(x).$$

Moreover, $\bar{\mu}$ is not necessarily unique.

Remark C.2 ($p > 1$). When $p > 1$, the proof follows in a similar manner. One will arrive at

$$a_j = \text{Mean}(\alpha_1^j, \dots, \alpha_M^j).$$

Proof. Assume that $N = 1$, so each measure μ_i is comprised of only one atom α_i for $i = 1, \dots, M$. WLOG we can also assume that the sequence $(\alpha_i)_{i=1}^M$ is non-decreasing. By convexity of the function $\phi_a(x) = |x - a|$ for $a \in \mathbb{R}$, the Wasserstein barycenter will also have $N = 1$ atoms. Then, by (21) the problem of finding the barycenter $\bar{\mu}$ is equivalent to the optimisation

$$(55) \quad \inf_{\nu \in \mathcal{P}_1(\mathbb{R})} \sum_{i=1}^M W_1(\mu_i, \nu) = \inf_{a \in \mathbb{R}} \sum_{i=1}^M |a - \alpha_i| = \inf_{a \in \mathbb{R}} \sum_{i=1}^M \phi_{\alpha_i}(a).$$

The minimiser $a^* \in \mathbb{R}$ to the right-hand side of (55) is obtained by solving $df/dx(x) = 0$ over \mathbb{R} , where

$$f(x) = |x - \alpha_1| + \dots + |x - \alpha_M| = \phi_{\alpha_1}(x) + \dots + \phi_{\alpha_M}(x).$$

Since

$$\frac{d\phi_{\alpha_i}}{dx}(x) = \text{sgn}(x - \alpha_i) \quad \text{for } i = 1, \dots, M,$$

we have that

$$(56) \quad a^* = \arg \inf_{a \in \mathbb{R}} \sum_{i=1}^M |a - \alpha_i| = \text{Median}(\alpha_1, \dots, \alpha_M).$$

In particular, if $M \bmod 1 = 0$, then $a^* \in [\alpha_{M/2}, \alpha_{M/2+1})$. If $M \bmod 2 = 1$, then the (unique) optimizer is given by $a^* = \alpha_K$ where $K = \lfloor M/2 \rfloor + 1$. Setting $a = a^*$ gives (54).

If $N > 1$, then the problem of finding the Wasserstein barycenter is equivalent to

$$\inf_{\nu \in \mathcal{P}_1(\mathbb{R})} \sum_{i=1}^M \mathcal{W}_1(\mu_i, \nu) = \inf_{(a_1, \dots, a_N) \in \mathbb{R}^N} \sum_{i=1}^M \sum_{j=1}^N |a_j - \alpha_i^j|.$$

Interchanging the order of summation, we see that

$$\inf_{(a_1, \dots, a_N) \in \mathbb{R}^N} \sum_{i=1}^M \sum_{j=1}^N |a_i - \alpha_i^j| = \sum_{j=1}^N \left(\inf_{a_j \in \mathbb{R}} \sum_{i=1}^M |a_j - \alpha_i^j| \right).$$

By applying (56) to each summation over M , we obtain the desired result (54). \square

We conclude this section with a formal statement of the algorithm associated to WK-means.

Algorithm 2: WK-means algorithm

Result: k centroids

calculate $\ell(r^S)$ given S ;

define family of empirical distributions $\mathcal{K} = \{\mu_j\}_{1 \leq j \leq M}$;

initialise centroids $\bar{\mu}_i, i = 1, \dots, k$ by sampling k times from \mathcal{K} ;

while $loss_function > tolerance$ **do**

foreach μ_j **do**

 | **assign** closest centroid wrt \mathcal{W}_p to cluster $\mathcal{C}_l, l = 1, \dots, k$;

end

update centroid i as the Wasserstein barycenter relative to \mathcal{C}_i ;

calculate $loss_function$;

end
