Minimum variance portfolios with enhanced stability

William Smyth [a]

[a]Accounting, Finance and Economics Department, Ulster University Business School, Cathedral Quarter, Belfast, Northern Ireland, United Kingdom.

ARTICLE HISTORY
Compiled July 20, 2023

ABSTRACT
We present a novel machine learning-driven approach to identify equities which may be used to build minimum variance portfolios that exhibit enhanced stability relative to an index portfolio. We assess the impact on both the clustering and stability monitoring processes of using correlation-based and entropy-based distance metrics. The interaction between entropy-based clustering outcomes and correlation-based clustering outcomes is nuanced. We investigate both and identify and explain the discrepancies. We find that the entropy-based distance metric has an edge based on practical considerations. We demonstrate how changes in stationarity of the underlying portfolio-specific excess returns data generating process are readily picked up by clustering and reflected in the resulting stability monitoring output. We discuss the optimal number of clusters. Our findings can assist in enhancing the stability of minimum variance portfolios.

KEYWORDS
Modern Portfolio Theory, Proximity matrix, Clustering, Entropy

1. Introduction

This paper presents a way to enhance the stability of minimum variance portfolios using entropy-based proximity matrices and clustering techniques. Estimating correlation matrices lies at the heart of many applications in quantitative finance. This is despite correlation having limitations when it comes to capturing the codependency of random variables. Correlation captures only linear codependence and is significantly impacted by outliers. The efficacy of its application is arguably limited to multivariate Gaussian processes which are not the norm in finance.

To address these potential shortcomings, some finance scholars have turned to the use of other mechanisms for evaluating codependence. Increasingly, concepts from information theory are becoming the subject of focus. In particular, the concept of Shannon entropy has particular relevance in finance. Entropy attempts to quantify the uncertainty associated with the outcome of a random variable. This has obvious appeal in machine learning in general and in financial machine learning in particular. In this study we therefore incorporate entropy as a construct for distance measurement and hence also for clustering. We compare the use of entropy-related measures with traditional correlation-based measures and in doing so produce results which add to the pool of knowledge surrounding the use of entropy in assessing codependency in quantitative finance applications.

CONTACT William Smyth, Author. Email: w.smyth@ulster.ac.uk

Minimum variance portfolios are subject to varying degrees of stability and robustness. Our results demonstrate that clustering based on an entropy-based distance metric such as normalised *variation of information* identifies clusters of index stocks which may be used to build a minimum variance portfolio that is more stable than one constructed from the entire index. Variation of information can be thought of as the information theory analogue of linear algebra's correlation. It has the added advantage that it can capture non-linear dependence. Maasoumi and Racine (2002) argue that information entropy is particularly useful in addressing failings of the dominant correlation measures of fit, predictability, and dependence. They argue this is particularly the case during periods on market non-linearity, such as market crashes.

We focus on minimum variance portfolios as they are less susceptible to estimation error. Gennotte (1986) observed that the asset prices required to construct the efficient frontier are not observable, and therefore must be estimated. These estimates are subject to error. Klein and Bawa (1976) illustrated the importance of this element in optimal portfolio choice. They point out that over time optimal asset weights are both noisy and unstable. Mean variance optimal portfolios rely on estimates of mean returns. The minimum variance portfolio does not (Merton (1980)). Lopez de Prado (2016) identifies the main sources of estimation error in using empirical correlation matrices as a basis for allocating component weights to minimum variance portfolios and introduces techniques to address the issue.

Our paper is structured in two parts. In the first part we address the general idea of clustering, an unsupervised machine learning technique used to group objects in the absence of labels. Clustering is underpinned by a proximity matrix which may be a similarity matrix or a dissimilarity matrix. The distinction is made contingent upon whether similarity or distance is being used to compare entities. In this study we focus exclusively on the use of distance metrics to define proximity. Thus, when we refer to a proximity matrix, a distance matrix is implied.

The objects which are subject to clustering are the stocks which make up market indices such as the S&P 400, the S&P 500, the S&P 600 and the Russell 1000. The proximity matrix is made up of elements which quantify distance between pairs of objects. This in turn implies the existence of features upon which comparisons can be made. In our study the observable data are excess daily returns for each stock collected over a defined period of time. Our daily returns data span the twenty-year period 2002 to 2021 inclusive. However, throughout the course of our analysis we variously focus on non-overlapping subsets of this time period. The rationale for doing so will be explained in-situ when this has taken place. We use daily 30-day US Treasury Bill prices to determine the benchmark risk-free rate of interest.

At the heart of defining distance is the notion of a distance metric and to that end the primary focus of our study is to compare the use of two distinct types of distance metric: metrics based on entropy; metrics based on correlation. By making use of the extensive data at our disposal (across indices, time-periods and stocks) we are able to explore in detail the role distance metric type plays in the clustering process.

In the second part of the paper, we view the clusters produced in the first part through the lens of portfolio stability. Again, the focus here is to assess how clusters perform depending on whether they were produced as a result of clustering on an entropy-based metric or a correlation-based metric. This has real practical implications for the selection of collections of stocks on which to build minimum variance portfolios. In particular we explore the impact of clustering on the stability of clusters (subsets) of an index relative to other clusters and to the index as a whole.

2. Background

Modern Portfolio Theory posits that the weights w of an optimal portfolio can be calculated from the covariance matrix W and vector α of its expected excess returns (Elton and Gruber (1997). The returns generating processes for securities in this portfolio are assumed to be stationary and Gaussian. The covariance matrix W is estimated from historical time series of returns. This makes the matrix noisy and as mentioned subject to error. The optimization process magnifies this noise. Portfolio optimization requires certainty in information input, however investment information in the real world is uncertain. The traditional way to approach the problem of estimation error is to increase the number of observations T (Ledoit and Wolf (2003)). A more recent approach is that of re-sampling proposed by Michaud (1989) whereby Monte Carlo simulation is used to compute statistically similar alternative portfolios.

Pafka and Kondor (2004) call estimation error the "curse of dimensions". The problem is that estimating a $N \times N$ correlation matrix using N time series each of length T, where T is bounded, inevitably introduces estimation error. For large N this can overwhelm to the extent that the justification for the applicability of the theory is called into question. Gennotte (1986) pointed out that such estimation error becomes an even bigger problem when the distribution is not observable.

The minimum variance portfolio is a set of N securities where the overall risk of the portfolio is minimized for a given performance target τ. According to Modern Portfolio Theory (MPT), it is located on the efficient frontier in risk-return space as depicted in Figure 1.

INSERT FIGURE 1 HERE

MPT dictates that the optimal minimum variance portfolio is a vector of weights $w := (w_1, \ldots, w_N)^*$, derived through a quadratic optimization process. Variance $<w, Cw>$ is minimized subject to the constraint $<w, g> \geq \mu$, with C being the population correlation matrix, g a vector of predictors and μ fixed:

$$w = \tau \frac{C^{-1} g}{g^* C^{-1} g} \tag{6}$$

The hierarchical structure of correlation in financial markets was investigated by Mantegna (1999), wherein he documents the hierarchical arrangement of traded securities.

Zhou, Cai, and Tong (2013) provide a review of the uses of entropy in finance. Similar to our research question, they did investigate its uses in Modern Portfolio Theory. They directed their attention to the Mean Variance portfolio rather than the minimum variance portfolio we focus on. Entropy has been used in the literature to measure risk in preference to a variance-based risk modeling approach (Philippatos and Wilson (1972), Yang and Qiu (2005), Simonelli (2005), Smimou, Bector, and Jacoby (2007)). Meanwhile, Chen, Xing, Xu, Zhao, and Principe (2016) gave some inter-disciplinary insights into minimizing the entropy of the estimation error whereby information can be preserved as much as possible.

In the econophysics literature several papers have quantified the degree of statistical uncertainty present in a correlation matrix (Drożdż, Kwapień, Grümmer, Ruf, and Speth (2001), Burda and Jurkiewicz (2004),and Pafka and Kondor (2003). These suggest that covariance matrices constructed from financial time series contain so much noise that their structure appears to be random. Pafka and Kondor (2003) argue that this is because the noise becomes stronger with an increase in portfolio size, until it reaches a point that over-rides the available information.

Xu, Zhou, and Wu (2011) develop a λ mean-hybrid entropy model to address the uncertainty present in the portfolio selection problem and investigated the ability of an entropy measure to the mean-variance-skewness of a portfolio. More recently, Smyth and Broby (2022) present a Stability Measure that builds on the seminal work of Marchenko & Pastur (1967a). They derived an analytic form for the probability density function of the eigenvalue distribution of a random covariance matrix. This is formed from a data matrix whose elements are independent identically distributed random variables drawn from a zero-mean process with finite variance.

We note that estimation error is less of a problem with minimum variance portfolios. Rao (1971) suggested that variance should be minimized to obtain the best unbiased estimators.

$$\Sigma_{shrink} = \alpha F + (1-\alpha)\Sigma_{SCM}, 0 \leq \alpha \leq 1 \tag{5}$$

where the convex combination Σ_{shrink} has a shrinkage target F and shrinkage estimator Σ_{SCM}, with α being the shrinkage intensity.

A common measure of estimation risk in minimum variance portfolio portfolios is given by TrE^{-1}/TrC^{-1} which is very close to unity when T is sufficiently large for a given N, (El Karoui (2010))

$$\frac{TrE^{-1}}{TrC^{-1}} = \frac{q}{q-1} \tag{7}$$

which holds for a wide class of processes. As can be seen, the out-of-sample risk TrE^{-1} can far exceed the true optimal risk TrC^{-1} when $q = T/N$ is not very large, diverging as $q \to 1$.

3. Distance Metrics

It is instructive to provide some detail in terms of the measures and metrics used in this paper. We begin with correlation-based measurement. The most common definition of correlation between two statistical variables, X and Y, is the Pearson correlation coefficient which we may represent as $\rho_{X,Y}$. The first point to note is that $\rho_{X,Y}$ is not a true metric. It has symmetry but does not meet either of the conditions of nonnegativity or the triangle inequality. However, it is possible to derive a metric from $\rho_{X,Y}$. We consider two such correlation-based metrics commonly used in the literature as part of our analysis. Namely,

$$CM_1 : d_\rho[X,Y] = \sqrt{(1-\rho_{X,Y})/2} \tag{1}$$

$$CM_2 : d_{|\rho|}[X,Y] = \sqrt{1-|\rho_{X,Y}|} \tag{2}$$

These metrics are similar in structure but have one important distinction. CM_1 treats positive and negative correlation differently. In fact, it considers negative correlation as being more distant than positive correlation, even if the absolute value of the correlation is the same. There are applications in finance where this is desirable. For instance, in building a long-only portfolio, holdings in negatively correlated components can only serve to offset risk, and thus need to be treated differently to positively correlated components for the purposes of diversification. On the other hand, in the case of long-short portfolios we would treat positively and negatively correlated components similarly since the position sign can override the sign of the correlation. This is the case in CM_2 where no distinction is made between positive and negative correlation.

Turning our attention to entropy-based measures, Sulthan and Jayakumar (2016) illustrate the use of entropy in finance to determine the likelihood of a specific behaviour by a security. For a theoretical summary see Rényi (1961). We commence with a definition of Shannon entropy. Consider a discrete random variable X which takes values from a set S_X with probability $P(x)$

$$H(X) = -\sum_{x \in S_X} P(x) \log P\left(x\right) \tag{3}$$

$\log 1/P\left(x\right)$ may be interpreted as the level of surprise associated with outcome $X = x$ and so the summation over all outcomes computes the expected level of surprise. Entropy, therefore, represents the amount of uncertainty associated with the random variable X.

With regard to measuring the codependency of random variables this definition needs to be extended to accommodate the notion of joint entropy where a second discrete random variable Y is included. Y takes values from the probability space S_Y with probability $P(y)$. The joint entropy of X and Y is given by

$$H(X,Y) = -\sum_{x \in S_X} \sum_{y \in S_Y} P(x,y) \log P\left(x,y\right) \tag{4}$$

It is important to note that entropy is finite only for discrete random variables. This must be taken into account when dealing with continuous random variables such as stock returns. In this paper we follow the approach of Jaynes (2003) in discretizing the sample space.

In order to produce an information theoretic counterpart to correlation we need to consider the concept of mutual information. Mutual information provides a measure of the informational gain in X resulting from knowledge of Y, and is defined as follows in terms of joint and marginal entropy

$$I\left(X,Y\right) = H\left(X\right) + H(Y) - H(X,Y) \tag{5}$$

which in explicit summation form reads

$$= \sum_{x \in S_X} \sum_{y \in S_Y} P(x,y) \log \left[\frac{P(x,y)}{P(x)P(y)}\right] \tag{6}$$

Mutual information has obvious intuitive appeal, but it is not a true metric. It satisfies the conditions of nonnegativity and symmetry but not the triangle equality. Thus, we must take one further conceptual step before arriving at an entropy-based distance metric. That step will take us from mutual information to variation of information, defined as

$$VI\left(X,Y\right) = H\left(X,Y\right) - I(X,Y) \tag{7}$$

Variation of information represents the expected uncertainty in one variable when we know the value of the other. It is a true metric but it does not have a formal upper bound because $H(X,Y)$ depends on the specifics of S_X and S_Y. This would preclude fair comparison between populations of different sizes and so, to produce a bounded metric we normalize to unity, confining the range to the closed interval [0,1]. The ranges of the correlation-based metrics defined above are also confined to this interval. Thus, we have an entropy-based distance metric in the form of the normalized variation of information $\widetilde{VI}\left(X,Y\right)$,

$$EM_1 : \widetilde{VI}\left(X,Y\right) = \frac{VI(X,Y)}{H(X,Y)} = 1 - \frac{I(X,Y)}{H(X,Y)} \tag{8}$$

As was the case for the correlation metrics we consider a second entropy-based metric, closely related to normalized variation of information. Kraskov, Stögbauer, and Grassberger (2011) incorporated the single modification of replacing $H(X,Y)$ with its lower bound $\max\left\{H\left(X\right), H(Y)\right\}$ to produce, in general, a more sensitive metric in

$$EM_2 : \widetilde{VI}_K\left(X,Y\right) = 1 - \frac{I(X,Y)}{\max\left\{H\left(X\right), H(Y)\right\}} \tag{9}$$

As outlined above these definitions apply to discrete random variables. Thus, when it comes to dealing with continuous random variables, as we do in this study in the form of stock returns, we must adapt our approach accordingly. In effect we course-grain the observations into discrete bins B_X and B_Y, and apply the concepts on the binned observations. The process of binning introduces an element of subjectivity and outcomes may be biased as a result. To address this, we incorporate the results of Hacine-Gharbi and Ravier (2018) and Hacine-Gharbi, Ravier, Harba, and Mohamadi (2012). These studies identify optimal binning arrangements for marginal and joint entropy respectively, as shown below. For marginal entropy

$$B_X = \text{round}\left[\frac{\gamma}{6} + \frac{2}{3\gamma} + \frac{1}{3}\right]$$

$$\gamma = \sqrt[3]{8 + 324N + 12\sqrt{36T + 729T^2}} \tag{10}$$

with T representing the number of observations. In the case of joint entropy

$$B_X = B_Y = \text{round}\left[\frac{1}{\sqrt{2}}\sqrt{1+\sqrt{1+\frac{24\text{T}}{1-\hat{\rho}^2}}}\right] \tag{11}$$

where $\hat{\rho}^2$ is the estimated correlation between X and Y.

Thus, we are now in a position to employ formal distance metrics to evaluate the dissimilarity (distance) between pairs of random variables. We have four metrics in total; two correlation-based; two entropy-based. This allows us to compare the impact of using different metrics of the same type in addition to metrics of different types. Using these metrics on pairs of random variables (pairs of series of stock returns) we can generate proximity matrices for entire indices. In this paper we perform this for four major indices: S&P 400; S&P 500; S&P 600; Russell 1000. It is on the resulting proximity matrices that clustering is performed. We provide further detail of this in the following section.

4. Clustering

Machine learning problems are typically categorised as being supervised or unsupervised in nature. The distinction is made around the presence or absence of labels. A label may be thought of as the manifestation of the interaction of a collection of features to produce an overall summary outcome specific to an observed entity (the object). For instance, in domestic property valuation the label may be the sale price of a house. The features are the set of quantifiable factors which contribute to the determination of the label value. By analogy with the domestic property market, features may take the form of the number of rooms in the dwelling, floor area, presence or absence of a garden, historic renovation costs, median local salary levels, pollution statistics, crime statistics, quality categorizations for local schools, proximity of industrial premises, accessibility to medical services, hospitality services, public amenities, shopping outlets, etc., and potentially dozens more. The true list of features is in general unknown, but an assumed list of features is engineered as part of the modelling process. Training and testing are guided (supervised) by labels when they are available but must proceed without guidance when labels are unavailable or are not relevant.

Clustering is an example of unsupervised machine learning. The objective is to utilise information implicit in the features to group objects into clusters, in the absence of an explicit label for each object. The allocation of objects to clusters is underpinned by the notion of similarity (conversely distance). It is an optimisation procedure designed to maximise intragroup similarity and minimise intergroup similarity. Conversely, yet equivalently, clustering aims to minimise intragroup distance and maximise intergroup distance. Clustering arises naturally in many areas of finance, particularly in portfolio management, and in risk management more generally.

The process begins with a data (or observation) matrix of size $N \times F$, with N representing the number of objects and F the number of features. A proximity matrix of size $N \times N$ is then generated, each element representing the proximity of two of the N objects. The proximity matrix may be a similarity matrix where the elements are based on a measure of similarity such as correlation or mutual information. Or it can be a distance matrix where proximity is a distance measure and again may be correlation-based or entropy-based. Ideally the proximity measure will also be a true metric satisfying the conditions of nonnegativity, symmetry and the triangle inequality (Kraskov et al. 2008).

Clustering may be partitional or hierarchical in nature depending on whether nesting is present. Partitional clustering produces a single level unnested partitioning of objects so that each object belongs to one and only one cluster. Hierarchical clustering, as the name suggests, produces a sequence of nested partitions, each one a sub-division of the previous. At the top level is a single cluster containing all the objects. Then, depending on the extent of the hierarchy, each level of partitioning contains more and smaller clusters. Taken to the extreme this may ultimately result in a collection of singleton clusters. Each individual object being its own cluster.

Additionally, depending on the precise definition of cluster, there are different types of clustering algorithms available. In this article we adopt the concept of centroid and utilize a k-means algorithm. k-means clustering is a powerful technique though it comes with two potential limitations; firstly, the number of clusters is user-specified and hence may not be optimal; secondly, the initialisation of cluster centres is random and hence the effectiveness of the algorithm may be compromised. We address both these considerations. The first consideration is resolved by alluding to one of the main objectives of our study; to demonstrate the impact of clustering on index portfolio stability. In Smyth and Broby (2022) we developed a technique to monitor the stability of a minimum variance index portfolio of stocks. One of the key results of this paper is the demonstration of increased stability in a minimum variance portfolio as a result of clustering. The process we developed to monitor stability depends on valid application of the Marchenko-Pastur theorem. As in the previous article we focus on a quality parameter value of $q = 1.5$, the ratio of rows to columns in the data matrix. One outworking of this is that we must observe a practical lower bond on the number of stocks in a cluster. The more clusters we allow, the smaller each cluster will tend to be. Additionally, there may be investment-related considerations such as concentration of risk and diversification limitations which will also have an input into the desirability of a particular cluster size. Ultimately, the optimal number of clusters in this study is dictated by technical considerations (Marchenko and Pastur (1967b)) and investment considerations (risk concentration) rather than alluding exclusively to a fundamental quality measure such as a silhouette coefficient (Rousseeuw (1987), Lopez de Prado (2019)) as may otherwise be the case. However, we monitor the silhouette coefficients for all clusters identified. As a result, we limit our consideration to clustering arrangements of size $k \in \{2, 3, 4, 5\}$. We find that our main research questions are fully answered within this scope.

The formula used to compute a silhouette coefficient is

$$S_i = \frac{b_i - a_i}{\max(a_i, b_i)}; \ i = 1, .., N \tag{12}$$

For a given clustering (i.e., for a given random initialization) S_i is a silhouette coefficient evaluated for each stock in the index. a_i represents the average distance between stock i and all other stocks in the same cluster. b_i represents the average distance between stock i and all elements in the nearest cluster of which stock i is not a member. In essence this is a comparison between intercluster distance and intracluster distance. S_i is constrained by $|S_i| \leq 1$. Values close to 1 reflect a stock that is well allocated whilst values closer to -1 reflect the opposite.

A single unitless figure for overall clustering quality is then given by

$$Q = \frac{E\left[\{S_i\}\right]}{\sqrt{Var\left[\{S_i\}\right]}} \tag{13}$$

Where $E\left[\{S_i\}\right]$ and $Var\left[\{S_i\}\right]$ are the mean and variance of the silhouette coefficients respectively. We use Q to distinguish from the Marchenko-Pastur quality parameter q. We employ this quality measure to address the second potential limitation of k-means clustering. In performing clustering for a given value of k, we permit a large number of random initializations and select the initialization with the largest silhouette quality value.

The data matrix which forms the starting point for our clustering analysis is based on historic daily excess returns values for the collection of stocks spanning an index such as the S&P 500. Thus, each column of the S&P 500 data matrix is a daily excess returns time series for that index stock. In order to generate the proximity matrix (in this paper a distance matrix), a distance metric is evaluated on the empirical data in each pair of columns, spanning all possible stock-pair combinations in the index. We perform this for the full range of distance metrics. Specifically, for the correlation-based metrics CM_1 and CM_2, and for the entropy-based metrics EM_1 and EM_2.

Once generated, clustering takes place on the proximity matrix elements. By design, objects and features are one and the same in this study. They are both a list of index stocks.

4.1. k-means Clustering: method

In this paper we investigate the impact of using entropy-based distance metrics over correlation-based distance metrics for the purposes of clustering. We also consider how distance metrics compare (entropy v correlation) in terms of how the clusters they predict perform under our stability monitoring process. The core data we work with is historic daily excess returns data for a range of market indices: S&P 400, S&P 500; S&P 600; Russell 1000. The historic data span the calendars years 2002 to 2021 inclusive. We break this twenty-year period of time into smaller time periods of interest. Specifically, we consider time periods identified in Smyth & Broby (2022) as being of significant interest from a stability point of view. Additionally, we also consider time periods coinciding with major global events such as the global financial crisis and the post-Covid19 period. By stability we refer to a period of time over which the signal-to-noise ratio of the empirical correlation matrix is stable. Equivalently, a time over which the portfolio-specific data generating process may be considered stationary. We employ all four of the distance metrics in our analysis. Specifically,

EM_1: $\widetilde{VI}$, normalized variation of information.

EM_2: $\widetilde{VI}_K$, normalized variation of information (Kraskov et al. (2008)).

CM_1: d_ρ, discriminating on the sign of the correlation.

CM_2: $d_{|\rho|}$, indifferent to the sign of the correlation.

The process involves applying a distance metric to a pairing of stocks within an index of N stocks, using the empirical excess returns for each stock over a pre-defined period of time in order to generate an element in a proximity matrix. By iterating this over all possible pairwise combinations we populate the $N \times N$ proximity matrix. Each matrix is specific to the metric used to determine its elements. For instance, we may work with excess returns data spanning the post-Covid19 period to generate a proximity matrix for each of the distance metrics above combined with each of the stock indices. Alternatively, we may work with excess returns data spanning the global financial crisis to generate a separate set of proximity matrices, again one for each metric-index combination. We repeat this process to build up a rich source of proximity matrices which are subsequently subject to clustering.

Clustering is carried out using the k-means clustering algorithm for clusterings of size $k \in \{2, 3, 4, 5\}$. We incorporate a large number of random initializations to optimise each clustering outcome. We confine our interest to clustering with a maximum of five clusters. The rationale being that, later in our study, when we subject the clusters to stability measurement, it is important that we do not use clusters so small as to invalidate the Marchenko-Pastur conditions at the heart of our stability monitoring procedure. Thus, in the following section, where we assess various features and properties of the clusters, unrelated to stability measurement, we only consider clustering arrangements which we can justify using later for stability measurement. We subsequently compare clustering outcomes across different metrics and time periods and clearly demonstrate the influence these factors have. Clustering outcomes change markedly depending on the time periods within which the excess daily returns data are located, and also depending upon which distance metric was used to generate the proximity matrix. We ensure that time periods are sufficiently long that sample size is not an issue; the shortest time-period used contains approximately 450 daily returns. For a fixed collection of index stocks and a fixed distance metric it is not surprising that there should be differences found in clustering outcomes where the data is taken from different time periods. However, it should be borne in mind that if the portfolio-specific excess returns data generating process is stationary across these periods then (allowing for sample effects) we would not except the clustering outcomes to be significantly different. However, as we demonstrate below, this is not the case. Clustering outcomes can vary markedly, well beyond anything that could be attributed to sample effects. For clarity, when we compare clusterings across two distinct time periods we are naturally referring to a comparison between clusterings with an identical number of clusters. We then use a simple democratic process to arbitrate on matching clusters between clusterings, and we define commonality between clusterings as the percentage of stocks which were clustered identically in both clusterings. This is illustrated in Figure 2 using a fictional clustering for an index comprising 500 stocks with $k = 3$

INSERT FIGURE 2 HERE

4.2. k-means Clustering: results

We present the results of our clustering analysis by considering two distinct scenarios.

In the first scenario we select two time-period pairings; Pairing A and Pairing B. The pairings are fixed for the duration of this scenario. The pairings were chosen because they generally resulted in clustering outcomes which differed considerably even for the same metric (Pairing A), or because they differed only marginally (Pairing B). For each time-period pairing we considered each index in turn, then each metric in turn, then each value of k, and recorded the resulting clustering commonality.

In a second scenario we fix the market index, and the time period, and consider how clustering arrangements differ within that time period depending on whether an entropy-based metric or correlation-based metric was used. We then repeat this across all time periods and record values for maximum and minimum agreement between metric types. This is recorded for all four clustering options for k. In order to present results concisely, we present results for only two distance metrics; EM_1 and CM_1. In other words, one entropy-based metric and one correlation-based metric.

For the vast majority of the cases we considered, the difference in clustering outcomes between the two correlation-based metrics was minimal (never less than 97% in common). There was slightly greater discrepancy (never less than 93% in common) between the two entropy-based metrics but the noteworthy distinction comes from comparing entropy-based and correlation-based metrics. Depending on the precise situation considered these metrics produced clustering outcomes which differed by as little as 1% (in other words they identically allocated 99% of the stocks in an index of typically 500 stocks) to differing by as much as 68% (meaning the identically allocated only 32% of the stocks in an index of typically 500 stocks).

When comparing results across different time periods it is the entropy-related metrics which proved more discerning, though this should not be interpreted as better clustering (we shall elaborate on the latter consideration when we come to pass the clusters through the stability monitoring process). For instance, for all four indices, across the two time periods of Pairing A, between which clustering outcomes clearly differed, they differed more for the entropy-based metric than for the correlation-based metric. Similarly, across the two time periods of Pairing B, between which clusterings were most alike, again the entropy-based metric indicated greater commonality than the correlation-based metric. Table 1 summarizes our findings

INSERT TABLE 1 HERE

Time-period Pairing A corresponds to performing separate clustering analyses on returns data taken from two distinct two-year periods chosen because the clustering outcomes differed maximally. Time-period Pairing B corresponds to performing separate clustering analyses on returns data taken from two distinct two-year periods chosen because the clustering outcomes differed minimally.

By way of illustration, the figure of 88% indicates that for the two clusterings performed across Pairing B time periods, having used the correlation-based distance metric to populate the distance matrices, the resulting clusterings had 88% of S&P 400 index stocks allocated identically. The figure of 41% in the table above indicates that for the two clusterings performed across Pairing A time periods, having used the entropy-based distance metric to populate the distance matrices, the resulting clusterings had 41% of S&P 400 index stocks allocated identically.

The level of discrepancy manifesting in clustering outcomes between time periods in Pairing A (maximally 68%, the complement of 32%), supports our finding in Smyth & Broby (2022), that something fundamental has changed in the underlying returns data generating process between the periods in question. However, the main result here is that whilst the change is being picked up by both categories of distance metrics, the effect is heightened when using the entropy-based distance metric. This provides evidence for the preferential use of entropy-related metrics where enhanced sensitivity to change could play an important role.

In the second scenario we make a more direct comparison between entropy-based and correlation-based distance metrics. Specifically, for a given index and a given period of time of interest, we directly compare the resulting clusterings depending on which metric was used to determine distance in the first place. Results are summarized in Table 2.

INSERT TABLE 2 HERE

For clarity, we explain the interpretation of the figures 89 and 56 (89% and 56%) highlighted in the table. Clustering (with four clusters) was performed on two distance matrices, one matrix generated by the entropy-based distance metric and the other by the correlation-based distance metric. Each distance matrix was based on identical daily returns data from the S&P 500, taken from a period of time of interest. This was repeated for the full range of periods of time of interest. Across these periods, the resulting clusterings (i.e., the entropy-based clustering and the correlation-based clustering) for the specific time-period of interest were compared and it transpired that the maximum level of agreement was 89% and the minimum level of agreement was 56%. Agreement indicates the percentage of index stocks which were identically clustered under entropy-based and correlation-based clustering.

It is clear from the table that under identical circumstances clustering resulting from the use of the entropy-based metric as opposed to the correlation-based metric may be very similar (as much as 95% of stocks identically allocated) or they may be very different (as little as 50% identically allocated). Across all indices there is a common trend; the range between max and min being smallest in the case of $k = 2$ clustering. The range is notably larger, though similar, for $k \in \{3, 4, 5\}$ clustering. There is also a modest decrease in maximum agreement levels as the number of clusters increase which is to be expected.

In the previous section we referred to having followed Rousseeuw (1987) in assessing the quality of clustering using the silhouette method. This method of quality assessment was in-built to our process in terms of choosing which clustering outcome to adopt across a large number of random initializations (ceteris paribus). The purpose of this is to mitigate problems associated with k-means clustering using a random initialization of cluster centres. Adopting this approach means that for a given index, a given metric, a given time period, and a given value of k, we were able to identify the highest quality clustering outcome.

As an extension of this it is instructive to draw comparison between the quality of clustering outcomes across situations which differ only in the metric used. In such circumstances, the silhouette method invariably rates clustering via the correlation-based metric more highly than clustering via the entropy-based metric. Thus, if a decision is to be made taking only this particular quality measure into account, clustering via the correlation-based metric would be the clear choice. However, in this study, as in quantitative finance analyses generally, clustering is a means to an end and not the end itself. We take the clustering outcomes and pass them to the stability monitoring process. The de facto quality of a cluster is ascertained ex post based on the stability of the associated minimum variance portfolio. This, as we shall see in the following section, provides a more pragmatic assessment, indeed a more nuanced account of the relative efficacy of clustering on the basis of entropy-based or correlation-based metrics.

5. Portfolio stability

Smyth and Broby (2022) introduce a technique to monitor the stability (over time) of a minimum variance portfolio. To be precise the stability of such a portfolio is assessed via the stability of the signal-to-noise ratio in the associated empirical correlation matrix. The signal-to-noise ratio at any given time is determined via a novel application of the Marchenko-Pastur theorem. The quality parameter $q = T/N$ is a key parameter in the process: N represents the number of stocks in the portfolio; T represents the length of the time series of daily excess returns present in the underlying data matrix. In our study we fix $q = 1.5$; a practically optimum value. The $T \times N$ excess returns data matrix is used to determine the associated $N \times N$ empirical correlation matrix at time t. Thus, the empirical correlation matrix at time t is based on historic returns dating back to $t - T$, equivalently $t - qN$. In our analysis we roll this matrix through time, one day at a time, and determine a daily value for the signal-to-noise ratio of the empirical correlation matrix. Monitoring this value on an ongoing basis equates to monitoring the stability of the associated minimum variance portfolio as outlined in more detail below.

The eigenvalue distribution of the empirical correlation matrix is separated into those eigenvalues associated with noise and those eigenvalues associated with signal. The probability distribution of the noise-associated eigenvalues is compared with the Marchenko-Pastur pdf. An optimisation process is implemented to maximise the similarity between the two via kernel density estimation on the discrete empirical distribution. This optimisation is performed by tuning the process variance parameter (σ^2), explicitly stated in the formula for the Marchenk-Pastur pdf, shown here

$$p(\lambda) = \begin{cases} q\frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{2\pi\lambda\sigma^2}, & \lambda \in [\lambda_-, \lambda_+] \\ 0, & \lambda \notin [\lambda_-, \lambda_+] \end{cases} \tag{9}$$

where $\lambda_\pm = \sigma^2 \left(1 \pm \sqrt{q}\right)^2$.

The resulting σ^2_{opt} may be interpreted as the signal-to-noise ratio of the empirical correlation matrix. This matrix lies at the heart of our stability monitoring process and indeed is the only entity requiring estimation in the process of forming a minimum variance portfolio. This is why our approach is ideally suited to the management of such portfolios.

Monitoring σ^2_{opt} over time amounts to monitoring the stability of the minimum variance portfolio. A period of time during which stability is demonstrated may be considered a proxy for a stationary portfolio-specific data generating process. Detecting breaks (or being alerted to potential breaks) in stationarity has clear implications for effective portfolio management. It also has implications for the use of backtesting over periods of time which straddle a break in stationarity. In particular, if the backtest is part of any analysis based on some statistical approach which assumes stationarity.

This paper extends the consideration of stability monitoring beyond the case where a portfolio already exists. We employ the unsupervised machine learning technique of clustering to identify a collection of stocks from which to build a minimum variance portfolio. The collection of stocks in question (the clusters) are subsets of the stocks which comprise an entire index such as the S&P 500. Our results clearly demonstrate how clustering can be used to identify subsets of stocks which are more stable than the index as a whole.

Additionally, this paper addresses questions pertaining to the efficacy of the metric underpinning the clustering process. The clustering algorithm employed is that of k-means clustering. A powerful partitioning clustering algorithm. This approach to clustering requires a proximity matrix on which to operate. The proximity matrix in turns requires some means to quantify similarity or distance. As indicated from the outset we focus on a distance (proximity) matrix. The matrix elements are computed subject to some definition of distance. This may be entropy-based or correlation-based.

The rationale for considering distinct distance metric types was outlined in the introduction. We alluded to potential shortcomings in the use of correlation to assess codependency in statistical variables. Such shortcomings have prompted researchers to look at alternatives for codependency evaluation and one such choice is entropy. Shannon entropy to be precise. Thus, in this paper we disaggregate our treatment of the use of clustering to identify clusters of stocks which are more stable than the index from which they were drawn. We additionally investigate whether the use of entropy-based metrics or correlation-based metrics has a bearing on the outcomes of this process. As our results in the following section will show, the answer to this question is not straightforward. The relative efficacy of entropy-based metrics over correlation-based metrics in clustering (and stability monitoring) is highly nuanced. Suffice to say that both approaches warrant a place in any financial machine learning engineering toolbox.

6. Results

6.1. The Power of Clustering

As indicated above the first objective of this part of the paper is to demonstrate the capacity of clustering to identify subsets of index stocks which are more stable than the parent index. Thus, initially, we focus less on whether the underlying metric is entropy-based or correlation-based and more on the end result. Also, from the point of view of incorporating useful visualisations to demonstrate effects, we initially confine our interest to $k = 2$ clustering. In this region of k-space the distinction between entropy-based and correlation-based clustering is less obvious in any event. Recall from Table 2. that when $k = 2$, agreement between the two types of distance metric is greater than for larger values of k.

In order to present our results in a concise and coherent way we present results only for the post-pandemic time period. This has the additional benefit of being the most recent of the time periods under consideration. In terms of the bigger picture, this section of the paper is less about examining differences across time periods and more about the capacity of clustering to improve stability. As indicated in the introduction, choice of time period plays an important role as there can be marked differences in clustering outcomes depending on when in time the returns data were sourced. However, the trends and effects reported here for post-pandemic returns data largely apply across other time periods provided those periods do not span a readily identifiable break in stationarity. For instance, this contrasts the use of returns data straddling Q1 2020 when the full impact of Covid19 hit global stock markets. Or data straddling the global financial crisis of 2008.

Returns from the latter period are used later in this section to illucidate the dangers of using, as part of the same analysis, returns from before and after a period of dramatic change in the data generating process. In the case of the global financial crisis this amounts essentially to a discontinuity. We do not need an event of this severity to constitute a break in stationarity but it serves as an effective illustration.

To begin we present Figure 3, corresponding to excess returns data from the post-pandemic period (May 2020 to Dec 2021) for the S&P 600. In this example it is the entropy-based distance metric which is used to populate the proximity matrix.

INSERT FIGURE 3 HERE

It is instructive to describe what is being presented in each of the frames. Frame (a) clearly demonstrates how effective clustering is in identifying clusters which are highly differentiated in terms of the signal-to-noise ratio in the associated empirical correlation matrix. We have three readily discernible distributions, one for each of two clusters, and one for the index as a whole. Each distribution depicts the distribution of signal-to-noise ratio values ($\nu \equiv \sigma^2_{opt}(t)$) over the period of time in question. The rolling optimisation was performed using a daily step size but results are presented on a weekly basis as this is a more practical time period for monitoring. Frames (b) to (d) depict the distribution of weekly percentage change ($\Delta\nu$) over the same period. Frame (b) shows the distribution of $\Delta\nu$ for the index as a whole, whereas frames (c) and (d) depict the same for each cluster. What is clear from frames (c) and (d) is the difference in width of the distribution for each cluster. It is the width of this distribution that provides a heuristic for stability. The frames show how the distribution for one cluster is wider than that of the index whilst for the other cluster it is narrower than the index. The visualisation is quite compelling but clearly it is necessary to be more explicit. To that end we use two complementing measures of dispersion to assess relative stability between the clusters and the index from which they came. The first measure is a cluster specific coefficient of variation for the distribution of ν

$$CV_c = \frac{\sigma_c(\nu)}{\mu_c(\nu)} \tag{14}$$

This unitless quantity allows us to make comparisons between distributions with different means as is the case for the distribution of ν for the clusters and the index.

The second measure we use is simply the standard deviation of the distribution of weekly percentage change $\sigma_c(\Delta\nu)$. Later when we present results for the best performing clusters (Table 3), we adopt the convention that for a cluster to be deemed to outperform the parent index it must outperform it in both these measures. The term best performing cluster is relevant because as we increase the value of k it is possible that clustering may identify more than one outperforming cluster within a single clustering. When an outperforming cluster is identified, Table 3 records two properties of that cluster. The first is the percentage reduction in the value of the coefficient of variation as we move from the index (CV_I) to the cluster (CV_c)

$$\Delta CV_c = \frac{CV_I - CV_c}{CV_I} \tag{15}$$

The second property is the size of the cluster, expressed as a percentage of the parent index.

$$n_c = \frac{N_c}{N_I} \tag{16}$$

with N_c, N_I representing the number of stocks in the cluster and the index respectively.

Specifically, for Cluster 1 in frame (a) of Figure 3, we have $\Delta CV_{c_1} = 36\%$ and $n_{c_1} = 79\%$. In other words, 2-means clustering on a S&P 600 post-pandemic proximity matrix, populated by an entropy-based distance metric, has identified a cluster comprising 79% of the index stocks which results in a 36% reduction in the coefficient of variation compared to the index as a whole. A reduction in the coefficient of variation amounts to a narrowing of the distribution of values of ν over the period of time in question. A narrowing of the ν distribution amounts to higher levels of stability (within the context of forming a minimum variance portfolio) in this collection of stocks than in the parent index from which they were drawn. Whilst not listed here, the identity of the stocks in the cluster are of course known.

It is instructive to illustrate how the outcome of a clustering scenario can differ markedly, even if the only change across the entire parameter space is the type of distance metric used. To that end, in Figure 4, we present the results of reproducing the above scenario identically, other than having replaced the entropy-based distance metric with the correlation-based distance metric.

INSERT FIGURE 4 HERE

A couple of things are immediately evident from this graphic. In frame (a) it is apparent that the clustering has produced clusters which are distinct in terms of the distribution of ν. However, from inspection of the other frames there is no obvious distributional narrowing either between the two clusters or between either cluster and the parent index. Specifically,

$$\Delta CV_{c_1} = -117\%,\ \Delta\sigma_{c_1}(\Delta\nu) = -45\%,\ [n_{c_1} = 52\%]$$

and

$$\Delta CV_{c_2} = 28\%,\ \Delta\sigma_{c_2}(\Delta\nu) = -30\%,\ [n_{c_2} = 48\%].$$

Cluster 1 fails in both measures (-117% and -45%). Cluster 2 outperforms the index in terms of a reduction in the coefficient of variation (28%) but underperforms in terms of producing a narrowing in the distribution of $\Delta\nu$ (-30%). In summary, both clusters underperform the parent index in terms of stability when we apply the pair-collective of measures in identify outperforming clusters.

This example readily illustrates the potential for differences between the metrics. The current example favours the entropy-based metric. However, it is simply an illustrative example. As can be seen in Table 3, the situation is reversed on occasions. Indeed, there is a certain symmetry in the results presented in Table 3 which makes it clear there is no obvious winner in a head-to-head between the two metric types.

In Figure 5 we present results to reinforce the efficacy of our approach. Rather than perform clustering on a proximity matrix based on any distance metric we simply split the S&P 600 index into two essentially random clusters based on alphabetical order of stock tickers.

INSERT FIGURE 5 HERE

It is clear from frame (a) that when these randomly selected clusters are passed through the stability monitoring process there is no clear distinction in terms of the distributions of ν. Certainly not of the magnitude that is evident in Figure 3 and

Figure 4. A small amount of separation is apparent but when we inspect the scale on the horizontal axis, we see it is separation over a narrow portion of the axis. What we are observing here, essentially, are sample artefacts that are an unavoidable part of any statistical sampling process. The other frames depict a situation in which neither cluster appears decidedly narrower than the other. Indeed, for both clusters, the width of the respective distributions of $\Delta\nu$ appear to be greater than that of the index.

The visual impressions are supported by the measures ΔCV_{c_i} and $\Delta\sigma_{c_i}(\Delta\nu)$,

$$\Delta CV_{c_1} = -48\%,\ \Delta\sigma_{c_1}(\Delta\nu) = -73\%,\ [n_{c_1} = 50\%] \tag{17}$$

and

$$\Delta CV_{c_2} = -106\%,\ \Delta\sigma_{c_2}(\Delta\nu) = -67\%,\ [n_{c_2} = 50\%]. \tag{18}$$

Both clusters underperform the index on both measures.

Finally, the distributions in frames (c) and (d) fail a two-sided two-sample Kolmogorov-Smirnov hypothesis test of samples coming from the same population (Massey Jr (1951)). In other words, we cannot reject the null hypothesis that they come from the same distribution. In contrast the distributions in frames (c) and (d) of Figure 3 and Figure 4 both pass this test with vanishing p-value. Collectively this provides a clear illustration of the value inherent in our approach in terms of its capacity to produce clusters with statistically significant differences in stability. And importantly, clusters which are more stable than the index of which they are part.

In Figure 6, we highlight the danger of using data which straddle a change in the underlying data generating process. The point is powerfully made by drawing on data which straddles the extreme event that was the global financial crisis of 2008. The data underpinning Figure 6 are drawn from a two-year period symmetrically straddling (one year pre- and one year post-) the crash.

INSERT FIGURE 6 HERE

Frame (a) demonstrates clear cluster separation. However, the distribution of ν for each cluster is bipolar in nature. In fact, more than bipolar, there is in essence two distinct distributions for each cluster. This is also the case for the index as a whole. We may readily conclude from this that the combination of k-means clustering and our stability monitoring process was highly effective at capturing the discontinuity that marked the crash. There is clear regime change in terms of the underlying data generating process and this reflected in the results. Moving to frame (b), the most obvious feature is the width of the horizontal scale required to capture the distribution. Additionally, the distribution for Cluster 1 has marked negative skew. This is in stark contrast to the equivalent frames of earlier figures and is a direct consequence of the large downward movements which comprised the crash event. We do not present values for ΔCV_{c_i} and $\Delta\sigma_{c_i}(\Delta\nu)$ for this situation. Consistent with our position in this paper we should not attempt to draw statistical inference from a situation where data has been used which clearly straddles a break in stationarity of a portfolio-specific data generating process.

6.2. Optimum Number of Clusters

Stability relates to the fluctuation over time of the signal-to-noise ratio of an empirical excess-returns correlation matrix. For the purposes of this paper, it is measured in two ways. The first way is through the evaluation of a coefficient of variation for the mean weekly value of ν; one value for the index as a whole (CV_I) and one for each individual cluster (CV_{c_i}). This is then expressed as a percentage decrease (ΔCV_{c_i}) as we move from index to cluster. Only the figures for the best performing cluster are recorded. The second way is through a basic measure of dispersion for the weekly percentage change ($\Delta\nu$) in ν. In this study, for a cluster to be reported as being more stable than the index it needs to outperform the index in both these measures. This raises the bar over using either of the measures in isolation. We performed clustering for the range of k-means options, $k \in \{2, 3, 4, 5\}$, in addition to the distance metric types of entropy-based (EM_1) and correlation-based (CM_1). For clarity, setting $k = 3$ means that the clustering algorithm is forced to select exactly three clusters, not at most three clusters.

Our analysis was carried out for four market indices; the S&P 400, the S&P 500, the S&P 600 and the Russell 1000. Results presented in Table 3 are based on excess returns collected over a two-year period post global coronavirus pandemic. For this period clustering based on EM_1 and CM_1 has the following commonality features: with $k = 2$, EM_1 and CM_1 produce clusters which have approximately 80%-90% of index stocks allocated identically, depending on the index; with $k \in \{3, 4, 5\}$, EM_1 and CM_1 produces clustering outcomes for which approximately 60-80% of index stocks identically allocated, again depending on the index.

Thus, we would naturally expect differences between entropy-based and correlation-based clustering to show up more as k is increased. Table 3 below records values for only one cluster within each clustering arrangement. That being the best performing cluster.

INSERT TABLE 3 HERE

We have already established that clustering, on the whole, readily identifies clusters which are more stable than the whole index of stocks from which the clusters were extracted. The focus here is to summarize outcomes across regions of parameter space which produce noteworthy results; specifically, across k-values and distance metric types. In terms of interpreting values recorded in the table, it is constructive to explain the values 21% and 46% in the first S&P 400 row. Firstly, these values relate to a clustering arrangement of $k = 2$, where the metric used to populate the proximity matrix was the entropy-based metric EM_1, the normalised variation of information. The value of 21% reflects the percentage reduction (ΔCV_{c_i}) in the value of the coefficient of variation as we move from the whole index (CV_I) to the individual cluster (CV_{c_i}). In each scenario, it is the value for the best performing cluster which has been reported. The value 46% reflects the size of the best performing cluster as percentage of the number of stocks in the wider index. Cells of the table populated by 0 indicate that none of the identified clusters were found to outperform the index in that particular region of parameter space. Inspecting the $k = 2$ section of the table we see that an outperforming cluster was found for each of the indices. In the cases of S&P 500 and S&P 600 only one of the two metric types was successful in finding an outperforming cluster. The general pattern is repeated as we move from $k = 2$ to $k = 3$; an outperforming cluster found for all indices, but not for all metrics and all indices. The best performing of the $k = 3$ clusters outperform their $k = 2$ counterparts and are smaller. This is true for both metric types. It is also apparent at this stage that for a given value of k, the correlation-based metric is identifying smaller clusters, in general, than the entropy-based metric.

Moving down the table to the arrangement where the clustering algorithm is set to $k = 4$, there is one significant observable change. For the S&P 400 and S&P 600 index data utilised, neither the entropy-based or the correlation-based metric, under the k-means clustering algorithm employed, are able to identify an outperforming cluster. What is also apparent is the continuing trend of outperforming clusters being both smaller and better performing than for smaller values of k. Thus, as we force the algorithm to split the index into more clusters it is proving more difficult to find an outperforming cluster. However, when an outperforming cluster is found, it is both better performing and smaller than for the $k = 2$ and $k = 3$ settings. A trend is also continuing in which the correlation-based metric is sourcing smaller clusters than the entropy-based metric.

Finally, in the $k = 5$ section of the table, we see that neither of the metric types were able to source outperforming clusters for the S&P 400, S&P 500 and S&P 600 index data. Both metric types were able to source an outperforming cluster for the Russell 1000 index data. Indeed, with the continuing trend of improved stability over lower k-values and smaller clusters. Yet again the correlation-based metric sourced a smaller cluster than the entropy-based metric.

We believe the explanation of the inability of either distance metric type to source an outperforming cluster for any of the S&P indices is linked to the absolute size of the clusters. In particular, if we look at the cluster size in the R1000 case for the correlation-based metric we see a value of 13%. Clusters of this size for the smaller S&P indices, which have typically only half the number of stocks of the Russell 1000, are stretching the validity of the Marchenko-Pastur theorem on which our stability analysis is based. There are implied restrictions on the size of the data matrix for which the Marchenko-Pasteur theorem can accurately predict the eigenvalue distribution for the associated correlation matrix. The theorem dictates eigenvalue distributions explicitly based on a quality parameter q which is the ratio of the number of rows (T) to the number of columns (N) in the returns data matrix. Firstly, q must always exceed 1. Throughout our analysis q has been fixed at 1.5 for all instances so this condition is met. However, underpinning the requirement that $q > 1$ are implications for the absolute values of the two components that form the ratio. Namely, N, the number of stocks in the index or cluster being analysed, and T, the length of the time series of returns. Of the two dimensions, N is the constraining one in the context of our analysis. N should be large, according to the theorem, and this means that as we start sourcing smaller clusters, we are inevitably working with smaller data matrices. In particular, we are working with data matrices where the number of columns is potentially too small for the theorem to be valid.

We have not formally tested this explanation in this paper as the key research questions we asked have been answered by considering clustering arrangements up to and including $k = 5$. As a consequence of these observations, we have not explored the parameter space for clustering arrangements beyond $k = 5$. Inspecting the table as whole it is not possible to say which of the entropy-based or correlation-based metrics are better at sourcing outperforming clusters. As is frequently the case for many applications in financial machine learning, we recommended the inclusion of both approaches in any analysis toolbox. There is some suggestion that on average the best performing clusters found by the entropy-based metric are better performing than their correlation-based counterparts. In parallel with this we can observe that the cluster sizes identified by the entropy-based metric are larger than those identified by the correlation-based metric. Cluster size has practical significance when taken in conjunction with other factors. This may give the entropy metric an edge in terms of the value-add. A supporting argument for this is incorporated into our conclusion.

7. Conclusion

In this paper we have demonstrated the effectiveness of the unsupervised machine learning technique, clustering, to identify clusters of stocks within an index which are more stable than the index as a whole. The outcomes are sensitive to the time period over which the data is collected. This sensitivity cannot be attributed to small sample effects. The problem arises because of fundamental changes in the underlying data generating process over time. This has obvious implications for performing statistical analyses using data taken across breaks in stationarity. In particular, if the statistical analysis being undertaken assumes stationarity in the data generating process. By combining k-means clustering with our stability monitoring process we are able to identify outperforming clusters and subsequently monitor the stability of a minimum variance portfolio formed from the stocks in that cluster.

Other areas of focus in our study are the distance metric type used to generate the proximity matrix and the number of clusters used. With regard to metric type, we considered both entropy-based and correlation-based metrics. Across indices, across time periods, and across the number of clusters used, we showed there may be strong agreement or strong disagreement in the clustering outcomes which result based on the metric type employed. Our analysis demonstrates the capacity of both type of distance metric to identify clusters which produce significant stability improvements over the parent index. The effects are sufficiently nuanced to mean it is advisable to utilise both metrics types in any given analysis, and choose between the two in a bespoke way. Two potentially significant observations were made vis-à-vis choice of metric type. The first was that when both metrics result in improvement, in the case of otherwise identically scenarios, the entropy-based metric produces slightly greater improvement than the correlation-based metric. The second is that, when the metrics identify outperforming clusters, the entropy-based metric tends to identify larger clusters than the correlation-based metric. This cluster size effect is also nuanced by the number of cluster centres used in the analysis. Both these factors should be considered in the round when deciding on an optimum number of clusters to use.

In order to inform the compound decision on which distance metric to use and what constitutes an optimum number of clusters, it is important to draw distinction between stability and other factors which would clearly play an important role. For instance, the identification of a stable collection of stocks on which to base the minimum variance portfolio is a separate matter to selecting a global minimum variance collection of stocks. There may be a trade-off between stability and global minimum variance and so an optimization process will be required to locate a universal best choice of cluster. Having identified a high stability subset of stocks there are other important considerations which will feed into a decision on cluster size. One, for instance, is around whether the selected subset constitutes an undesirable concentration of risk, perhaps through under diversification. There is also the technical consideration that since our stability monitoring process depends on the validity of the Marhcenko-Pastur theorem, clusters cannot be too small otherwise results may be erroneous. A benchmark for precision on what constitutes too small is not the subject of this study.

Suffice to say it was straightforward to recognise when certain clusters were clearly too small based on erroneous output. All clusters presented in our results section have been identified as valid for the purposes of Marchenko-Pastur application. Use of a fundamental cluster quality metric such as the silhouette coefficient can also feature in the decision process as an additional input. However, practical considerations will trump any such clustering quality measure when arriving at a final decision. In this article, clustering is a means to an end, not the end itself.

In summary, we have presented a turn-key process for enhanced implementation of minimum variance index portfolio investment. Machine learning has been effectively deployed in identifying clusters of index stocks with which to build minimum variance portfolios with enhanced stability dynamics relative to the index from which they were taken. We have identified a range of factors which should be taken into account as part of this process and provided a rationale for how the impact of each factor can be assessed. Collectively what we have presented constitutes a powerful mechanism for the improved selection and management of minimum variance portfolios.

Figure 1

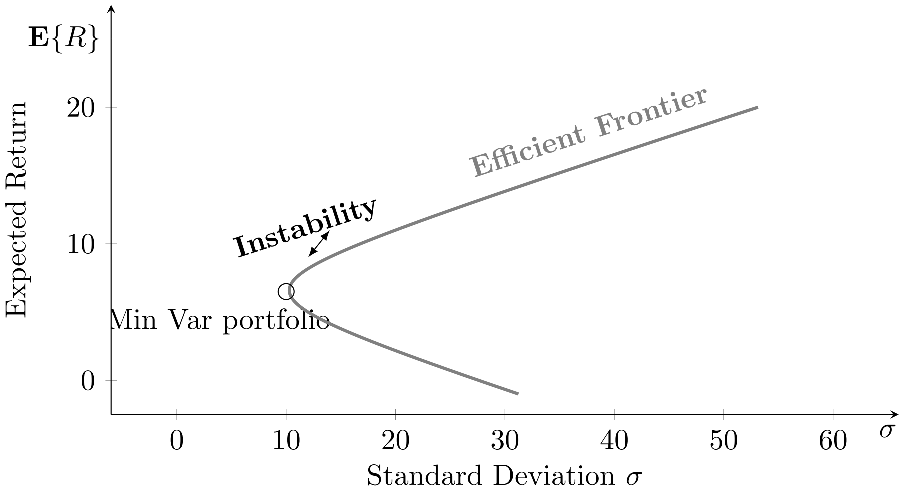

Figure 1. An efficient frontier in risk-return space with the positioning of the minimum variance portfolio. It also highlights that there is instability in the positioning of the frontier and that portfolio due to estimation error.

Figure 2

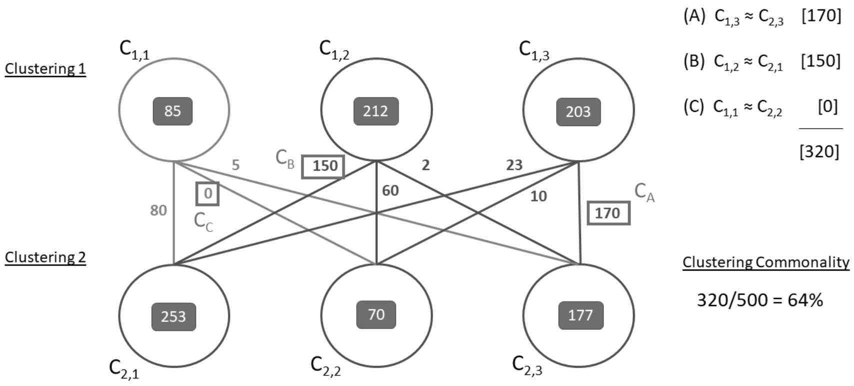

Figure 2. Determining commonality between distinct clustering outcomes for a fictional index comprising 500 stocks with $k = 3$

Figure 3

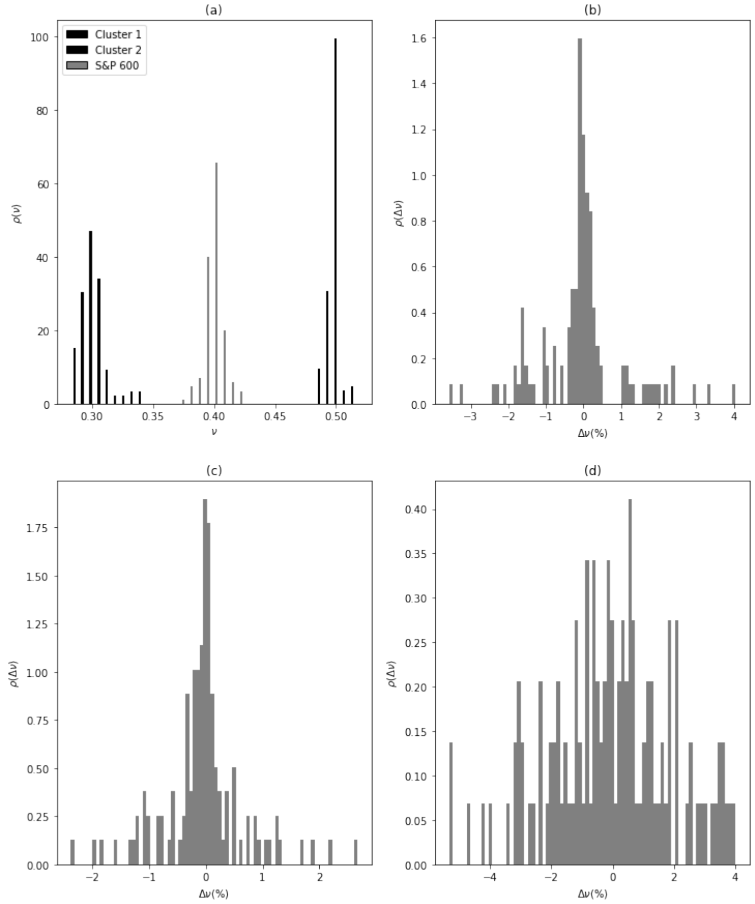

Figure 3. Entropy-based clustering: (a) Distribution of ν for Cluster 1, Cluster 2 and the whole index. (b)-(d) Distribution of weekly percentage change in ν for the entire index (b), for Cluster 1 (c), and for Cluster 2 (d).

Figure 4

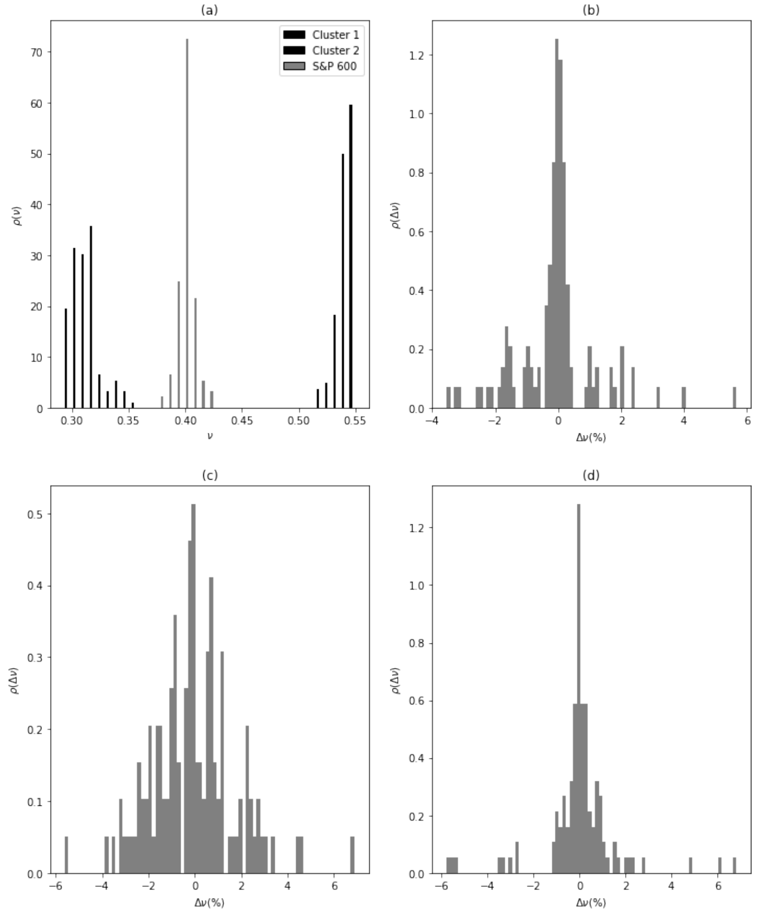

Figure 4. Correlation-based clustering: (a) Distribution of ν for Cluster 1, Cluster 2 and the whole index. (b)-(d) Distribution of weekly percentage change in ν for the entire index (b), for Cluster 1 (c), and for Cluster 2 (d).

Figure 5

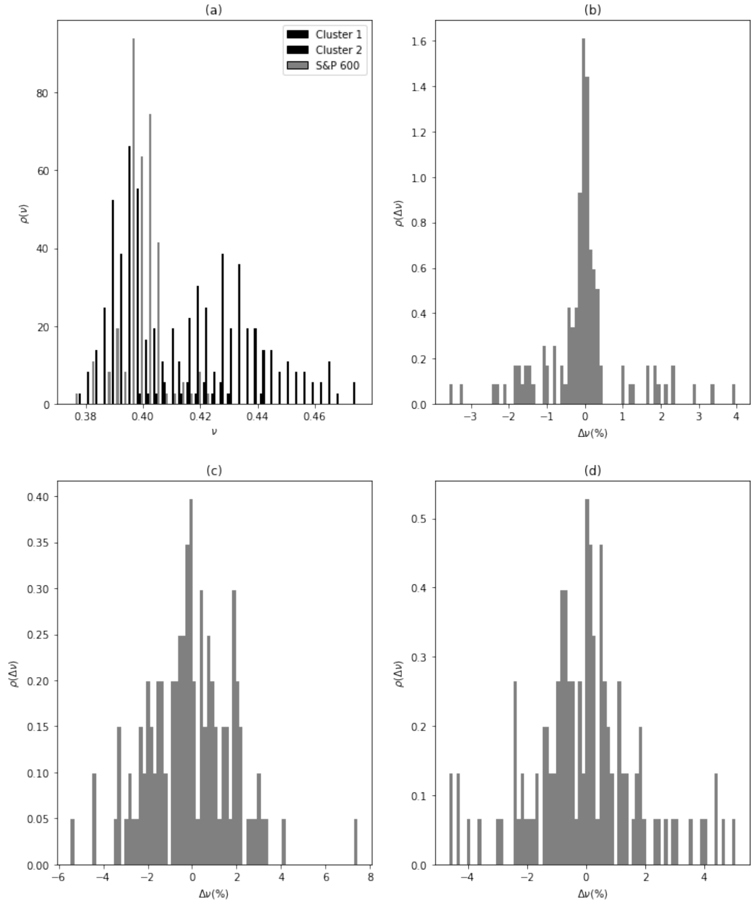

Figure 5. Random clusters based on alphabetical order of stock tickers: (a) Distribution of ν for Cluster 1, Cluster 2 and the whole index. (b)-(d) Distribution of weekly percentage change in ν for the entire index (b), for Cluster 1 (c), and for Cluster 2 (d).

Figure 6

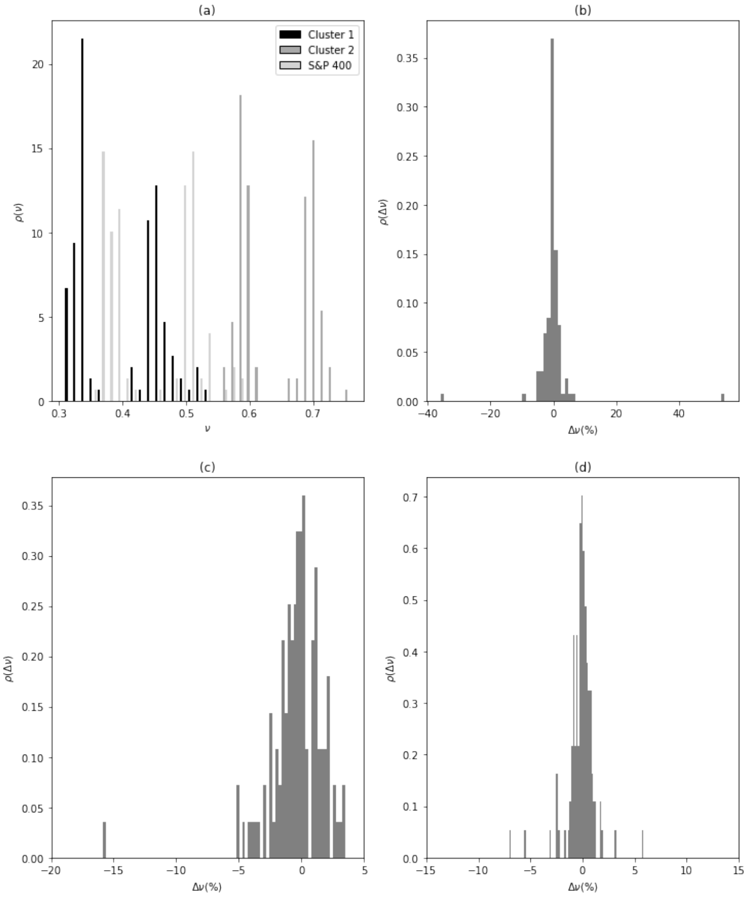

Figure 6. (a) Distribution of ν for Cluster 1, Cluster 2 and the whole index. (b)-(d) Distribution of weekly percentage change in ν for the entire index (b), for Cluster 1 (c), and for Cluster 2 (d).

Table 1

Market Index	Stocks (%)	
	EM_1	CM_1
Time-period Pairing A		
S&P 400	**41**	45
S&P 500	32	35
S&P 600	32	34
Russell 1000	32	32
Time-period Pairing B		
S&P 400	99	**88**
S&P 500	91	85
S&P 600	89	79
Russell 1000	89	81

Table 1. Sensitivity of clustering outcomes to distance metric type across time-period pairings.

Table 2

Index	k-means							
	2		3		4		5	
	max	min	max	min	max	min	max	min
S&P 400	94	62	94	51	90	58	86	56
S&P 500	94	83	90	51	**89**	**56**	89	50
S&P 600	95	81	89	54	89	50	82	50
Russell 1000	93	82	86	57	86	58	86	51

Table 2. Maximum and minimum levels of commonality in clustering outcomes dependent on distance metric type, across indices and number of clusters.

Table 3

Index	EM_1		CM_1	
	a	b	a	b
	$k = 2$			
S&P 400	**21**	**46**	26	49
S&P 500	0	-	17	48
S&P 600	36	79	0	-
Russell 1000	23	54	31	43
	$k = 3$			
S&P 400	38	32	0	-
S&P 500	31	46	32	22
S&P 600	0	-	39	22
Russell 1000	37	35	32	37
	$k = 4$			
S&P 400	0	-	0	-
S&P 500	41	29	40	21
S&P 600	0	-	0	-
Russell 1000	38	30	32	16
	$k = 5$			
S&P 400	0	-	0	-
S&P 500	0	-	0	-
S&P 600	0	-	0	-
Russell 1000	49	30	45	13

Table 3. Clusters outperforming the index on stability measures, across indices, distance metric type and number of clusters. (a) $\Delta CV_c(\%)$ (b) $n_c(\%)$.

References

Burda, Z., & Jurkiewicz, J. (2004). Signal and noise in financial correlation matrices. *Physica A: Statistical Mechanics and its Applications*, *344*(1-2), 67–72.

Chen, B., Xing, L., Xu, B., Zhao, H., & Principe, J. C. (2016). Insights into the robustness of minimum error entropy estimation. *IEEE transactions on neural networks and learning systems*, *29*(3), 731–737.

Drożdż, S., Kwapień, J., Grümmer, F., Ruf, F., & Speth, J. (2001). Quantifying the dynamics of financial correlations. *Physica A: Statistical Mechanics and its Applications*, *299*(1-2), 144–153.

El Karoui, N. (2010). High-dimensionality effects in the markowitz problem and other quadratic programs with linear constraints: Risk underestimation. *The Annals of Statistics*, *38*(6), 3487–3566.

Elton, E. J., & Gruber, M. J. (1997). Modern portfolio theory, 1950 to date. *Journal of banking & finance*, *21*(11-12), 1743–1759.

Gennotte, G. (1986). Optimal portfolio choice under incomplete information. *The Journal of Finance*, *41*(3), 733–746.

Hacine-Gharbi, A., & Ravier, P. (2018). A binning formula of bi-histogram for joint entropy estimation using mean square error minimization. *Pattern Recognition Letters*, *101*, 21–28.

Hacine-Gharbi, A., Ravier, P., Harba, R., & Mohamadi, T. (2012). Low bias histogram-based estimation of mutual information for feature selection. *Pattern Recognition Letters*, *33*(10), 1302–1308.

Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge university press.

Klein, R. W., & Bawa, V. S. (1976). The effect of estimation risk on optimal portfolio choice. *Journal of financial economics*, *3*(3), 215–231.

Kraskov, A., Stögbauer, H., & Grassberger, P. (2011). Erratum: estimating mutual information [phys. rev. e 69, 066138 (2004)]. *Physical Review E*, *83*(1), 019903.

Ledoit, O., & Wolf, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of empirical finance*, *10*(5), 603–621.

Lopez de Prado, M. (2016). A robust estimator of the efficient frontier. *Available at SSRN 3469961*.

Lopez de Prado, M. (2019). *Financial learning for asset managers*. Cambridge University Press.

Maasoumi, E., & Racine, J. (2002). Entropy and predictability of stock market returns. *Journal of Econometrics*, *107*(1-2), 291–312.

Mantegna, R. N. (1999). Hierarchical structure in financial markets. *The European Physical Journal B-Condensed Matter and Complex Systems*, *11*(1), 193–197.

Marchenko, V. A., & Pastur, L. A. (1967a). Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, *1*(4), 457.

Marchenko, V. A., & Pastur, L. A. (1967b). Distribution of eigenvalues for some sets of random matrices. *Matematicheskii Sbornik*, *114*(4), 507–536.

Massey Jr, F. J. (1951). The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, *46*(253), 68–78.

Merton, R. C. (1980). On estimating the expected return on the market: An exploratory investigation. *Journal of financial economics*, *8*(4), 323–361.

Michaud, R. O. (1989). The markowitz optimization enigma: Is 'optimized'optimal? *Financial analysts journal*, *45*(1), 31–42.

Pafka, S., & Kondor, I. (2003). Noisy covariance matrices and portfolio optimization ii. *Physica A: Statistical Mechanics and its Applications*, *319*, 487–494.

Pafka, S., & Kondor, I. (2004). Estimated correlation matrices and portfolio optimization. *Physica A: Statistical Mechanics and Its Applications*, *343*, 623–634.

Philippatos, G. C., & Wilson, C. J. (1972). Entropy, market risk, and the selection of efficient portfolios. *Applied Economics*, *4*(3), 209–220.

Rao, C. R. (1971). Minimum variance quadratic unbiased estimation of variance components. *Journal of Multivariate Analysis*, *1*(4), 445–456.

Rényi, A. (1961). On measures of entropy and information. In *Proceedings of the fourth berkeley symposium on mathematical statistics and probability, volume 1: Contributions to the theory of statistics* (Vol. 4, pp. 547–562).

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.

Simonelli, M. R. (2005). Indeterminacy in portfolio selection. *European Journal of Operational Research*, *163*(1), 170–176.

Smimou, K., Bector, C., & Jacoby, G. (2007). A subjective assessment of approximate probabilities with a portfolio application. *Research in International Business and Finance*, *21*(2), 134–160.

Smyth, W., & Broby, D. (2022). An eigenvalue distribution derived "stability measure" for evaluating minimum variance portfolios. *UNDER REVIEW - Quantitative Finance*.

Sulthan, A., & Jayakumar, G. D. S. (2016). On the review and application of entropy in finance. *International Journal of Business Insights and Transformation*, *10*(1), 14–18.

Xu, J., Zhou, X., & Wu, D. D. (2011). Portfolio selection using λ mean and hybrid entropy. *Annals of operations research*, *185*(1), 213–229.

Yang, J., & Qiu, W. (2005). A measure of risk and a decision-making model based on expected utility and entropy. *European Journal of Operational Research*, *164*(3), 792–799.

Zhou, R., Cai, R., & Tong, G. (2013). Applications of entropy in finance: A review. *Entropy*, *15*(11), 4909–4931.