Guide to Using Partition Clustering to Construct Portfolios

Gabriel Rambanapasi[a]

[a] *Stellenbosch University, Stellenbosch, South Africa*

Keywords: K-Means, Clustering, Price Momentun, Volatility, Diversification

1. Introduction

- aim to test momentum via cumulative return
- use clusters to group high performers, low performers to each group.
- build a portfolio, and check out returns.
- compare the underlying thesis with what has happended

2. Clustering and Appliactions to Asset Management

Unsupervised machine learning is a type of machine learning that searches for patterns in datasets with no pre-existing labels and a minimum of human intervention. One way in which unsupervised learning can be applied in data science and other quantitive disciplines is through clustering algorithms. Clustering is the process of grouping objects based on similar characteristics. The algorithms designed to cluster, achieve this function by connecting observation through distances, density of data points, graphs, or various statistical distributions. For a cluster to have meaning an algorithm has to maximize intra-cluster similarity and minimize inter-cluster similarity, such that each cluster contains information that's as dissimilar to other clusters(Kassambara, 2017). There exists various forms of cluster algorithms, each that addresses a broader task of analysis. The algorithms can be divided into two main types being partitioning clustering and hierarchical clustering. The major difference between the divisions of clustering is the partition clustering aims to specify a predetermined number of clusters whilst does not (Kassambara, 2017). Within partition clustering, for data with a small

*Corresponding author: Gabriel Rambanapasi
Email address: rambanapasi44@gmail.com (Gabriel Rambanapasi)

set of variables, K-means clustering and partitioning around medoids (PAM) are the most frequently used due to their fast compuation and simplicity. With K-means, each cluster is represented by the center or means of the data points belonging to the entire dataset. This makes the algorithm sensitive to outliers. However with PAM, each cluster is represented by one of the objects in the cluster. The other partition clustering algorithm used for datasets with a large number of variables is Clustering Large Applications (CLARA).

In asset management, key to funds generating superior risk adjusted returns is efficient portfolio diversification, thus presenting a great application for partition clustering. Stocks would be separated into groups through a clustering algorithm to maximize similarity within groups and minimizes similarity between groups. Thus allowing managers to select handpick stock to construct a diversified portfolio. Marvin (2015) use fundamental ratios (turnover and profitability ratios) weighted equally and K-means clustering to group US technology stocks listed on the NASDAQ and NYSE. A diversified portfolio is then constructed based on within cluster stock performance i.e. stock selected are those that possess the highest Sharpe ratio. Results over a period of 15 years that included the dot com bubble and the global financial crises showed that cluster portfolios exhibited more volatility than the benchmark (S&P 500), however returns to investors were above the benchmark at multiples ranging from 3.5 to 5.7 times when earnings are reinvested into the cluster portfolios. Bin (2020) uses a similar approach to Marvin (2015), however employing a combination of market ratios and fundamental ratios (price to earnings ratio, return on assets ratio and asset turnover ratio). From this study, compared to the S&P 500, portfolios constructed using market ratios under performed those that used fundamental ratios.

3. Data and Methodology

This section describes how we obtain the data set used in the study, details the clustering process and validating metrics employed, to obtain the results in 4.

3.1. Obtaining and prepariing the dataset

The data employed in this paper is based on the constituent list of the Johannesburg All Share Index (ALSI) from January 1, 2000 to January 15 2024, which contains the list of the 164 companies with the highest market value and liquidity. The historical price and volume data is retrieved from Yahoo Finance and fundamental data from Bloomberg. From historical price data obtained from Yahoo Finance, we filter stock that have trading volume that exceed 1 000 000 shares traded per year and exclude stock that have less than 90 percent of observations in the historical price dataset.

To avoid large oscillations in the data, we transformed the price series to include end of month data points thus returns calculations are based on from the monthly data. Monthly historical prices are transformed using simple returns and we assume that embedded in the price action are cooperate events such as stock splits or consolidations of the shares. Therefore there is no need to make additional transformations on the return series to reflect corporate actions.

The measures of similarity used in this study are volatility and price momentum. To cluster stock based on the two measures, we apply a percentile ranking criterion on stock scores during a time period. For price momentum, describes the causuality between relatively strong performance and high future return and vice versa. Ranking highly implies that strong performers and thus higher returns than weak performers. This study defines cross sectional momentum as the trailing 6 month cumulative return Jegadeesh & Titman (1993). For volatility, using a 12 month lockback period compute the standard deviation. The results of the ranking are shown in Table 6.1

3.2. Stocks clustering

3.3. Lloyd's algorithm

We employ the K-means that partitions n observations into k clusters ???. The goal is to minimize the within cluster sum of squares or analytically:

$$\arg\min_{s} \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|^2$$

where x are the observations, $S = S_1, S_2, \ldots, S_k$ are the sets of observations, and μ_i is the mean of the points in S_i. To arrive at the optimum number of clusters, we utilize the most popular algorithm

called the Lloyd's algorithm that is closely followed by Marvin (2015), Bin (2020) & Xu, Xu & Wunsch (2010).

Analytically:

Given a set of points $\{x_1, \cdots x_n\}\,(x_i \in \mathbb{R}^m)$,

- Initialize the K clusters with $\{C_1, \cdots C_K\}$ with centers $\{m_1, m_2, \cdots m_K\}\,(m_i \in \mathbb{R}^m)$. The centres are picked using the silhoutte method discussed in 3.4
- For all points $x_i (i \in \{1, \cdots n\})$, find the centre that closest based on a eucliean distance d. Following this, assign x_i to the cluster corresponding to the closest centre.

$x_i \in C_j$ if $d\,(x_i, m_j) \leq d\,(x_i, m_l) \quad (\forall l \in \{1, \cdots K\})(j \neq l)(\forall i \in \{1, \cdots n\})$.

- Recalculate the center for each cluster $C_l (l \in \{1, \cdots K\})$. The new cluster centres are the mean of the points in the same cluster.

$m_l = \frac{1}{|C_l|} \sum_{x_p \in C_l} x_p \quad (\forall l \in \{1, \cdots K\})$.

- Repeat processes two and three until no cluster has any change in point assignment.

3.4. Silhoutte index

To evaluate the goodness of fit of partitioning using K means clustering the silhoutte index is used. A summary that closely follows Rousseeuw (1987), can be seen below:

Given n data points $\{x_1, \cdots x_n\}$, a partitioning result of K cluster $\{C_1, \cdots C_K\}$ and distance metric d, for each x_i in cluster C_l, define

$a\,(x_i) = \frac{1}{|C_l|-1} \sum_{\forall x_j \in C_l, i \neq j} d\,(x_i, x_j)$

where $a(x_i)$ is the mean dissimilarity between x_i to all other points within the same cluster.

For each point x_i in cluster C_l, define

$b\,(x_i) = \min_{\forall p \in \{1, \cdots K\}, p \neq l} \frac{1}{|C_p|} \sum d\,(x_i, x_j\,)$

$b(x_i)$ is the minimum dissimilarity between x_i and all points in some C_p which does not contain x_i.

For each point x_i in cluster C_l, their silhoutte index is defined as

$$s\left(x_i\right) = \begin{cases} 1 - \frac{a(x_i)}{b(x_i)} & \text{if } a\left(x_i\right) < b\left(x_i\right) \\ 0 & \text{if } a\left(x_i\right) = b\left(x_i\right) \\ \frac{b(x_i)}{a(x_i)} - 1 & \text{if } a\left(x_i\right) > b\left(x_i\right) \end{cases}$$

where $s(x_i)$ ranges between $[-1, 1]$.

For $s(x_i)$ that approaches 1, it means that $a(x_i)$ needs to be significantly smaller than $b(x_i)$, implying that within-cluster mean dissimilarity is much less than the smallest between-cluster mean dissimilarity, and thus the model does a good job clustering similar points together.

For the $s(x_i)$ that approaches 0, $a(x_i)$ needs to be significantly greater than $b(x_i)$, implying that within-cluster mean dissimilarity is much greater than the smallest between- cluster mean dissimilarity, and thus the model does a poor job clustering similar points together.

For this study, we choose K with the highest silhoutte index/value

3.5. Portfolio Backtest

The out-of-sample performance of cluster portfolios is compared to the benchmark, the JSE All Share Index. To manage risk exposure to a single asset or industry, we use a cap on each asset's allocation. Thus use a single company methodology similar to Standard & Poor Capping Methodology [1]. In a single company capping methodology, no company in an index (cluster in our case) is allowed to breach a certain pre-determined weight as of each rebalancing period [2]. Theoretically, this should preserve the within cluster diversification benefits and allow the portfolio value to either increase or decrease depending on stock performance during the quarter. We rebalance the portfolios once every three months, similar to the frequency of rebalancing conducted by the JSE on the JSE/FTSE indices, that is, re balance on the last daty of March, June, September and December.

[1] see S&P (2023) for a breakdown of the process used to cap an index

[2] The maximum set for each stock in each cluster is its equal weight for that cluster i.e if there were 10 stock in cluster 1, the maximum weight would be 10%

4. Results

5. References

Asness, C. 2011. Momentum in japan: The exception that proves the rule. *The Journal of Portfolio Management*. 37(4):67–75.

Bin, S. 2020. K-means stock clustering analysis based on historical price movements and financial ratios.

Jegadeesh, N. & Titman, S. 1993. Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of finance*. 48(1):65–91.

Kassambara, A. 2017. *Practical guide to cluster analysis in r: Unsupervised machine learning*. Vol. 1. Sthda.

Marvin, K. 2015. Creating diversified portfolios using cluster analysis. *Princeton University*.

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.

S&P. 2023. *Index mathematics methodology*.

Xu, R., Xu, J. & Wunsch, D.C. 2010. Clustering with differential evolution particle swarm optimization. In IEEE *IEEE congress on evolutionary computation*. 1–8.

6. Appendix

6.1. Appendix A

Ticker	Volatility Rank	Price Momentum Rank
ABG	26.67	21.67
ANG	16.67	16.67
BIK	80.00	73.33
EQU	90.00	86.67
FFA	3.33	53.33
GFI	50.00	30.00
GLN	56.67	36.67
HAR	46.67	100.00
HMN	66.67	60.00
KAP	70.00	33.33
KBO	100.00	3.33
MCG	73.33	43.33
MRP	10.00	46.67
MTM	60.00	76.67
NPH	96.67	26.67
ORN	30.00	10.00
OUT	20.00	56.67
PAN	33.33	83.33
PIK	36.67	6.67
PPC	83.33	95.00
QLT	53.33	90.00
RMH	86.67	95.00
SAP	43.33	80.00
SBK	13.33	70.00
SOL	40.00	13.33
TCP	93.33	63.33
TFG	23.33	40.00
TRU	6.67	50.00
VKE	76.67	66.67
WHL	63.33	21.67

Table 6.1: Clustering Similarity Measures

6.2. Appendix B