NASDAQ Stock Portfolio Clustering Analysis

Ashton Koop

Building portfolios is all about diversifying risk across many different investments. In order for proper diversification, you want to build a portfolio of both low risk and high risk stocks in proportions that satisfy the investors risk preference. The use of clustering provides very useful tools for grouping stocks that have similar qualities together, so that a portfolio manager can pick and choose from different clusters in order to match their clients goals.

The clustering approach I used will be described below, and in each step will reference why this would be perfect for ensuring a portfolio manager can make the correct decisions for their clients.

Get correlation distances between stocks: The first step in the clustering process is to create a correlation matrix between stock returns and then convert this into distances.

The distances we compute can be interpreted as pairwise distances with quantity depending on the correlation between stocks. The exact formula used is

$$d_{ij} = \sqrt{2(1 - \rho_{ij})}$$

Since correlation () is a value between -1 and 1, calculating distance this way allows the structure to be as such

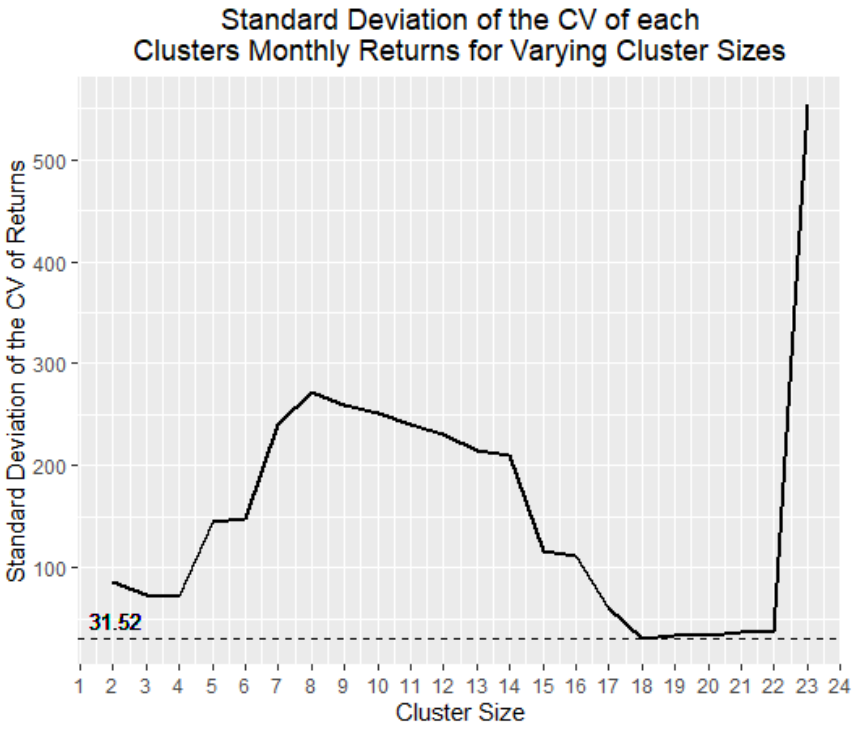
If $\rho_{ij} = 1$ then $d_{ij} = 0$ (perfect correlation, zero distance)
If $\rho_{ij} = 0$ then $d_{ij} = \sqrt{2}$ (uncorrelated, moderate distance)
If $\rho_{ij} = -1$ then $d_{ij} = 2$ (Perfect opposites, maximum distance)

Hierarchical Clustering with Complete Linkage: Imagine each cluster is a group of friends, and distance measures how badly they get along. You combine two friend groups and check all pairs; you then ask “How badly does the worst pair in one circle get along with the worst pair in the other”. You only let these friend groups merge if the worst pair of friends is better than the worst pair of friends in other potential groupings. By doing this, you find a grouping where the two people who get along the worst, still get along better than other combinations.

This is essentially what hierarchical clustering with complete linkage does to your data. It will only cluster groups with the smallest distance in the furthest pair of data points. This method ensures that data is grouped as tightly as possible. This is useful for the use case (portfolios that are highly correlated) as this groups highly correlated stocks as tight as possible.

$$Cluster\ Option\ = MIN_d\{MAX(d_{within\ cluster\ candidate})\}$$

Cutting the Tree and Selecting k = 18: In the R code, I looked at the standard deviation of the CV (coefficient of variation) of each cluster's monthly returns for varying cluster sizes. The CV metric is commonly used as a volatility measure, but can be interpreted as a unitless value measuring risk adjusted return. If the standard deviation of the CV is minimized, you are ensuring the cluster size makes each cluster as correlated (lowest variation of volatility) as possible. As seen in the figure, this metric is minimized at a cluster size of 18.



$$CV = \frac{\sigma_{monthly\ returns}}{\mu_{monthly\ returns}}$$

	StockSymbol	CV	Cluster
	<chr>	<dbl>	<int>
1	SPAN	4.11	1
2	NATH	4.63	2
3	STRA	3.40	3
4	AFAM	3.46	4
5	ACGL	3.63	5
6	AMED	3.79	6
7	GMCR	4.93	7
8	ITRI	5.33	8
9	SRCL	3.50	9
10	PRSP	4.19	10
11	NAFC	5.71	11
12	BRLI	5.75	12
13	FCZA	4.30	13
14	DECK	3.84	14
15	ISRL	4.95	15
16	NWSB	5.68	16

In order to build a portfolio to minimize overall risk, you need to pick and choose stocks that have shown to be low risk (volatility) in the past. The way I approached this is by selecting the lowest CV stock from each cluster, and grouping them all together into one portfolio. This allows me to build a well diversified portfolio (one from each cluster), that also minimizes overall risk (low CV)

If you invest equal amounts of money in each stock within this portfolio, your annualized return for the time period 2000-01-01 to 2009-12-31 would have been 24.29%. This seems high, however since the method was to pick the lowest CV stocks (historical data),

this essentially ensures high returns. Using historical data is not always a good predictor of future outcomes, so you must be careful not to interpret this as yielding ~24% forever onwards. However, since all I had to use was historical data, this is an optimal portfolio for minimizing risk.

To expand on this idea, it would be essential to allow forecasting stock returns, in order to accurately predict the future CV ratios for stocks and stock portfolio options.