3D Subspace Clustering for Value Investing

Kelvin Sim, *Institute for Infocomm Research*

Vivekanand Gopalkrishnan, *Deloitte Analytics Institute Asia*

Clifton Phua, *SAS*

Gao Cong, *Nanyang Technological University*

A 3D-subspace-clustering method generates rules to pick potential undervalued stocks; 3D subspace clustering is effective in handling high-dimensional financial data and is adaptive to new data.

Value investing is an investment strategy where the investor believes that a stock's fundamentals determine future stock prices.[1] A value investor analyzes stock fundamentals and buys stocks that are undervalued with the belief that the prices of the stocks will rise in the future.[2] The success of value investing is evident in the stock market, with many famous value investors' portfolios, such as Warren Buffett's, outperforming the market indices. There are also many successful mutual funds that follow the philosophy of value investing, such as Third Avenue Value Fund, which managed 5.04 billion dollars of assets in 2011. Academic research has shown that stock fundamentals are related to stock prices.[3]

Stock fundamentals can be measured by stock financial ratios. For example, the return-on-equity ratio measures a stock's efficiency in using its assets to generate profit, the debt–equity ratio measures the amount of the stock's assets that are debts, and the price–earnings ratio measures the ratio of the stock's current price to its current earnings.

Therefore, scrutinizing financial ratios is important in finding undervalued stocks.[2] However, there's no perfect rule that shows which financial ratios and what values of these ratios are related to undervalued stocks. For example, Benjamin Graham, the founder of value investment, prefers stocks with a price–earnings ratio of no more than 7.[2]

Using Graham's rules on picking stocks has been proven to generate profits for value investors. An experiment conducted over an eight year period from 1973 to 1980 showed this strategy to be profitable.[1] We propose using 3D subspace clustering to generate rules to pick potential undervalued stocks. The 3D subspace-clustering method is effective in handling high-dimensional financial data and is adaptive to new data. In addition, its results aren't influenced by human biases and emotions, and are easily interpretable. We conducted extensive experimentation in the stock market over a period of 28 years (from 1980 to 2007), and we found that using rules

1541-1672/14/$31.00 © 2014 IEEE
Published by the IEEE Computer Society

generated by two 3D subspace-clustering algorithms (CATSeeker and MIC) results in 60 percent more profits than using Graham's rules alone.

Problem Defined

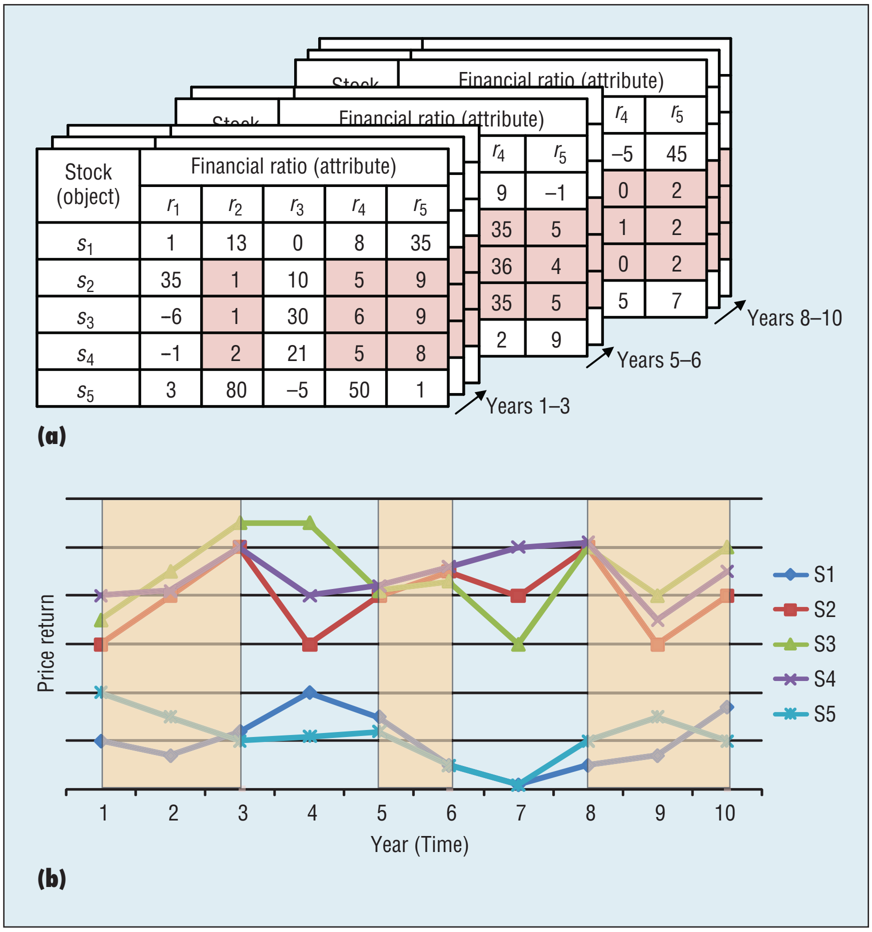
Value investing isn't simply about buying stocks based on some rules on the financial ratios, although Henry Oppenheimer[1] has shown that Graham's rule-based strategy enables the investor to make profits from the stock market. It's not necessary to always use Graham's rules, and the investor can set his or her own rules, based on his or her preferences and domain knowledge. These rules are generally used to select a comfortable number of stocks for the investor to conduct further analysis. Hence, these rules provide some general decision support.

For an inexperienced investor, manually setting rules on the financial ratios can be difficult, and even for the experienced investor, he or she might be prone to set irrational and biased rules. The investor can stick to Graham's rules, but the relevance of these rules at present time remains to be seen. Hence, the following problem needs to be addressed: How do we find rules on financial ratios that are related to high stock price returns? We should note here that we define the price return of a stock as (sold price – purchased price)/purchased price.

There are financial studies that investigate the impact of single financial ratios on stock prices.[3] However, different financial ratios quantify different aspects of a stock, so to get the complete picture, it will be useful to study the collective influence of financial ratios on the stock prices, and this is a nontrivial problem.

Proposed Solution

We propose using 3D subspace-clustering algorithms to mine rules that are related to high stock price returns.



Stock (object)	Financial ratio (attribute)				
	r_1	r_2	r_3	r_4	r_5
s_1	1	13	0	8	35
s_2	35	1	10	5	9
s_3	–6	1	30	6	9
s_4	–1	2	21	5	8
s_5	3	80	–5	50	1

Years 1–3

Stock	Financial ratio (attribute)	
	r_4	r_5
	9	–1
	35	5
	36	4
	35	5
	2	9

Years 5–6

Stock	Financial ratio (attribute)	
	r_4	r_5
	–5	45
	0	2
	1	2
	0	2
	5	7

Years 8–10

Figure 1. (a) Example of a 3D financial dataset defined by stocks, financial ratios, and years. The highlighted region is an actionable 3D subspace cluster of stocks s_2, s_3, s_4 that have similar financial fundamentals reflected in financial ratios r_2, r_3, r_4. (b) The price returns of the stocks. Stocks s_2, s_3, s_4 have high price returns.

The 3D subspace-clustering approach groups stocks that have similar fundamentals (financial ratios) and high price returns across years. The highlighted region in Figure 1a is a 3D subspace cluster containing stocks s_2, s_3, s_4 that have similar fundamentals reflected in financial ratios r_2, r_3, r_4 for years 1–3, 5–6, and 8–10. From Figure 1b, we can see that stocks s_2, s_3, s_4 have high price returns.

This cluster's subspace can be used as a rule that's related to high price returns. For future years, if there's a stock whose financial ratio values fall in this subspace, we can consider this stock as a potential undervalued stock. Using the example in Figure 1a, the rule is $r_2[2, 3]$, $r_3[10, 11]$, $r_4[5, 6]$ in year 1, …, $r_2[\ldots]$, $r_3[\ldots]$, $r_4[0, 2]$ in year 10, where $r[j, k]$ denotes that the stock's value on financial ratio r should fall between values j and k. If there's a stock whose set of future years contain this rule, this stock is a potentially undervalued stock.

3D subspace clustering is suitable for this value investing problem, due to the following reasons:

- *Effective in handling financial ratio data.* Financial ratio data is high dimensional, as the number of financial ratios and timestamps can be large. Techniques such as traditional clustering (for example, k-means clustering) suffer from the curse of dimensionality in this type of data; the stocks are equidistant from each other in the full space of the data, hence it's difficult to cluster them.[4] We developed 3D subspace clustering to overcome this curse of dimensionality. We achieve this by clustering stocks based on similar subsets of financial ratios (data subspace). The financial ratio data is continuous and 3D subspace clustering is generally used on this type of data.

- *Adaptive to new data.* The financial ratio data is constantly changing and 3D subspace clustering can be easily reapplied on the new data to get the updated results.
- *Easy interpretation of results.* The investor can easily analyze 3D subspace clusters because the clusters are explicitly created.

We aren't trying to solve the problem of how to invest, for example, or what stocks to buy at a particular time. Instead, we're trying to determine if 3D subspace clustering can help the value investor in his or her stock selection process by decreasing the pool of stocks to select.

To evaluate the effectiveness of using 3D subspace clustering for value investing, we compare its profits and risks to those of Graham's rule-based strategy.

Preliminaries

Let the 3D financial dataset be a cuboid $\mathcal{D}$ with its axes defined by objects (stocks) $\mathcal{O}$, attributes (financial ratios) $\mathcal{A}$, and time stamps (years) $\mathcal{T}$, for example $\mathcal{D} = \mathcal{O} \times \mathcal{A} \times \mathcal{T}$.

Let the value of financial ratio a of stock o, in year t, be denoted as v_{oat}. Let $p(o, t)$ be the closing price of stock o at the end of the fiscal year t, which we use as the buying and selling prices in our experiments. The price return of stock o, bought at year t and sold at year $t + i$, is calculated as

$$\text{ret}(o,t,t+1) = \frac{p(o,t+i) - p(o,t)}{p(o,t)},$$

which is also known as the i-period simple net return.[5]

For the sake of brevity, we also denote $\text{ret}(o)$ as the price return of stock o, if the year it's bought and the year it's sold aren't required to be explicitly stated.

A 3D subspace cluster is a subcuboid $C = O \times \mathcal{A} \times \mathcal{T}$, with its axes defined by a subset of stocks $O \subseteq \mathcal{O}$, a subset of financial ratios $A \subseteq \mathcal{A}$, and a subset of years $T \subseteq \mathcal{T}$. We denote $\{C_1, \ldots, C_m\}$ as the set of 3D subspace clusters mined from the dataset $\mathcal{D}$.

Research Design

We present the research design of our experiments. The research design consists of three main phases: data preparation, stock picking, and data analysis.

Data Preparation

In the data preparation phase, we obtained raw financial figures of US stocks from Compustat (see www.compustat.com) and converted this information into a 3D dataset of financial ratios. We removed microcap stocks (whose prices are less than \$5) from the data, as these stocks have a high risk of being manipulated and their financial figures are less transparent.

We converted the raw financial figures into 30 financial ratios, based on the ratios' formula from Investopedia (see www.investopedia.com/university/ratios).

We prepared a financial dataset $\mathcal{D}$ containing 30 financial ratios and spanning 28 years (from 1980 to 2007). The number of stocks increased from 3,335 to 5,049, due to the stock market's expansion. Some (14.7 percent) of the dataset contain missing values.

Graham's rule-based strategy uses 10 years of financial ratio data to pick stocks and to have a fair comparison; we also partition the financial ratio data into 10-year datasets, and use them as training data for the 3D-subspace-clustering algorithms to pick stocks. The partitioned datasets are denoted as $\mathcal{D}^t$, $t \in \{1980, \ldots, 1999\}$, with each $\mathcal{D}^t$ containing data of the set of years $T = \{t, \ldots, t + 9\}$. For example, $\mathcal{D}^{1980}$ contains data from 1980 to 1989.

We also processed these 10-year datasets $\mathcal{D}^t$ to contain only stocks that have high price returns. These datasets are required as training data for certain 3D-subspace-clustering algorithms. More specifically, $\mathcal{D}^t_{\text{min}_{\text{ret}}}$ is a processed dataset that contains stocks o, whose $CAGR(o, t, t + 9)$[3] min_{ret}, given that min_{ret} is a threshold. The compound annual growth rate is

$$CAGR(o,t,t+9) = \left(\frac{p(o,t+9)}{p(o,t)}\right)^{\frac{1}{9}} - 1.$$

We use compound annual growth rate instead of average return to minimize the effect of volatility of periodic returns.

We vary min_{ret} from 0.1 to 0.5, as there are no valid stocks in some 10-year datasets $D^t_{\text{min}_{\text{ret}}}$ for $\text{min}_{\text{ret}} > 0.5$.

Stock Picking

Graham's rule-based strategy consists of a buy phase and a sell phase.[1] In the buy phase, a stock is bought if it satisfies at least one reward criterion and one risk criterion. The criteria are shown in Table 1. If a stock satisfies at least one reward criterion and one risk criterion on year t, we will purchase the stock on the last day of its fiscal year t.

In the sell phase, the stock will be sold either on the last day of its fiscal year $t + 2$, or on the day when its price appreciates by 50 percent, whichever comes first. We slightly tweak the sell phase, as we are only able to obtain the price of the last day of each fiscal year of a stock. A stock will be sold on the last day of its fiscal year $t + 1$ if its price appreciates to more than 50 percent, or it will

Table 1. Graham's rule-based strategy (adapted from other work[1]).

Reward Criteria	
1	Earnings/price yield ≥ 2 × AAA bond credit rating yield
2	Price–earnings ratio ≤ 0.4 × highest price–earnings ratio of the stock during the past 5 years
3	Dividend yield ≥ 2/3 × AAA bond yield
4	Stock price ≤ 2/3 × tangible book value per share
5	Stock price ≤ net current asset value
Risk Criteria	
6	Total liabilities ≤ book value
7	Current ratio > 2
8	Total liabilities < 2 × net current asset value
9	Earnings growth of prior 10 years ≥ 7 percent annual (compound) rate
10	No more than two declines of 5 percent or more in year-end earnings in the prior 10 years
Financial Ratio Definitions in Reward and Risk Criteria	
Book value	Total assets to total liabilities
Current ratio	Current assets/current liabilities
Dividend yield	Dividend per share/stock price per share
Earnings/price yield	Earnings per share/price per share
Net current asset value	Current assets to total liabilities
Price–earnings ratio	Stock price per share/earnings per share
Tangible book value per share	Total tangible assets/total number of shares outstanding

be sold on the last day of its fiscal year $t + 2$. In our experiments, we show that this strategy still generates good profits.

We use 10 datasets $\mathcal{D}^t$, $t \in \{1989, \ldots, 1998\}$ as the testing datasets. Hence, the testing period contains 19 years, from 1989 to 2007.

For the 3D subspace-clustering strategy, depending on which algorithm we use, the training dataset can either be $\mathcal{D}^t$ or $\mathcal{D}^t_{\text{min}_{\text{ret}}}$. Assume that we mine 3D subspace clusters from $\mathcal{D}^t$, and we use the clusters as rules to pick stocks. Let $C = O \times A \times T$ be one of the 3D subspace clusters.

The general idea is to use the values of the financial ratios of a 3D subspace cluster as a rule to pick stocks, because these values are associated with stocks of high compound annual growth rate.

Let there be a 3D subspace cluster $C = O \times A \times T$ mined from training dataset $\mathcal{D}^t$, and let $V_{at_i} = \left\{V_{oat_i} \middle| o \in O\right\}$ be the set of values in cluster C, of financial ratio a at year t_i. We denote boundary(V_{ati}) = [min(V_{ati}), max(V_{ati})], which defines the boundaries of the values of financial ratio a at year t_i.

Definition 1 of our approach is the rule of the 3D subspace cluster where $C = O \times A \times T)$. We denote the rule of the cluster C as $\text{rule}(C) = \left\{\text{boundary}(V_{at_i}) \middle| a \in A, t_i, \in T\right\}$, which is a set of boundaries.

We then use rule(C) from training dataset $\mathcal{D}^t$ to pick stocks from dataset $\mathcal{D}^{t+9}$, which is the corresponding testing dataset. Subsequently, we use rule(C) from training dataset $\mathcal{D}^{t+j}$ to pick stocks from testing dataset $\mathcal{D}^{t+j+9}$, for $j \geq 1$.

Definition 2 of our approach is the 3D subspace-clustering strategy's buy rule. Let $\mathcal{D}^{t+9}$ be a dataset with a set of years $\mathcal{T}'$. A stock o' in dataset $\mathcal{D}^{t+9}$ is bought if it satisfies rule(C). Stock o' satisfies $\text{rule}(C) = \left\{\text{boundary}(V_{at_i}) \middle| a \in A, t_i, \in T = \left\{t_1, \ldots, \text{t}_{|T|}\right\}\right\}$ if, $\forall a \in A : v_{o'at'_1} \in \text{boundary}\left(V_{at_1}\right), \ldots, v_{o'at'_n} \in \text{boundary}\left(V_{at_n}\right)$ where $t'_n\ t'_1 \ldots t'_n \in \mathcal{T}'$, and $t'_1 < t'_2 < \cdots < t'_n$. The stock o' is then bought on the year t'_n.

We only buy stock o' at year t'_n once, regardless the number of times it's picked; this is to prevent the stock from dominating the results.

We consider 3D subspace clusters that contain at least two years, because clusters that contain only a single year are trivial. This means that the earliest year a stock can be bought in the testing data is in the second year. Hence, we set the testing data to start at year $t + 9$, so that the earliest year a stock in the testing data $\mathcal{D}^{t+9}$ can be bought is on year $t + 10$. If the testing data starts at year $t + 10$, then it's not possible to buy stocks on year $t + 10$ across all experiments.

To have a fair comparison, we used the same sell phase described for Graham's rule-based strategy in the previous "Stock Picking Phase" section for the 3D subspace-clustering strategy.

For a training dataset $\mathcal{D}^t$ or $D^t_{\text{min}_{\text{ret}}}$, we test the rules mined on the testing dataset $\mathcal{D}^{t+9}$. We use 10 datasets $\mathcal{D}^t$, $t \in \{1980, \ldots, 1989\}$ as the training datasets. Hence, each training dataset has a corresponding testing dataset. We use 10 datasets $\mathcal{D}^{t'}$, $t' \in \{1989, \ldots, 1998\}$ as the testing datasets.

Data Analysis

Let strat denote the strategy used in the stock picking phase. Let $O^{\mathcal{D}}{}_{strat}$ be the set of stocks bought using strat on a training dataset $\mathcal{D}$. We use the function ret to calculate the price return of the stocks bought.

Definition 3 is the average return of strategy:

$$ret^{\mathcal{D}}_{\text{strat}} = \frac{\sum_{o \in O^{\mathcal{D}}_{\text{strat}}} \text{ret}(o)}{|\, O^{\mathcal{D}}_{\text{strat}} \,|}.$$

The strategy's risk on training dataset $\mathcal{D}$ is its standard deviation of the average return. A high standard deviation implies that the strategy is risky and volatile. Let δ_{ret} denote the risk-free return that the investor is assumed to have. In calculating the standard deviation, we shouldn't incorporate returns that have at least δ_{ret}. Thus, we calculate the strategy's risk on training dataset $\mathcal{D}$ using the downside standard deviation, which is Definition 4 (the risk of strategy):

$$\text{risk}^{\mathcal{D}}_{\text{strat}} = \sqrt{\frac{\sum_{o \in O^{\mathcal{D}}_{\text{strat}} \mid \text{ret}(o) < \delta_{\text{ret}}} (\text{ret}(o) - \text{ret}_{\text{strat}})^2}{\mid \{o \mid o \in O^{\mathcal{D}}_{\text{strat}} \wedge \text{ret}(o) < \delta_{\text{ret}}\} \mid -1}}.$$

A strategy is thus desirable if it gives high average return and low downside risk (standard deviation), which can be measured using the Sortino ratio.[6] Definition 5 is the Sortino ratio of strategy:

$$\text{SortinoRatio}^{\mathcal{D}}_{\text{strat}} = \frac{\text{ret}^{\mathcal{D}}_{\text{strat}} - \delta_{\text{ret}}}{\text{risk}^{\mathcal{D}}_{\text{strat}}}.$$

We conduct different stock-picking strategies and evaluate their results by the following experiments:

- *Average returns across years.* We denote $\mathcal{T}$ as the set of years used to test a strat. For each testing dataset, we calculate $\mathcal{D}^t$, $t \in \mathcal{T}$, the average return $\text{ret}^{\mathcal{D}}_{\text{strat}}$, and $\text{SortinoRatio}^{\mathcal{D}}_{\text{strat}}$ of the stocks bought.
- *Overall average returns and risks.* We calculate the overall average return of a strat by averaging $\text{ret}^{\mathcal{D}t}_{\text{strat}}$, $\forall\, t \in \mathcal{T}$, and we calculate the overall risk of this strategy by averaging the downside standard deviation of $\text{ret}^{\mathcal{D}tt}_{\text{strat}}$, $\forall\, t \in \mathcal{T}$. We also calculate the overall Sortino ratios of the strategies using the overall average return and risk.

3D Subspace-Clustering Algorithms

We use a wide range of 3D subspace-clustering algorithms in our experiments.

TRICLUSTER

TRICLUSTER is the pioneer algorithm for mining 3D subspace clusters, which are denoted as triclusters.[7] A tricluster can be transformed into a wide variation of 3D subspace clusters, depending on the setting of the TRICLUSTER algorithm's parameters.[7] In a tricluster $C = O \times A \times T$, the stocks O have homogeneous values in the set of financial ratios A in each year $t \in T$, and the homogeneity and size criterion are satisfied subject to the setting of parameters δ, $\min_O$, $\min_A$, $\min_T$. We also set its parameters $\delta_y = \delta_z = \infty$, as they aren't applicable in mining our desired clusters. The clusters are sensitive to the parameters, and careful setting of the parameters is required.

STATPC

Moise and Sander proposed statistical significant subspace clusters (SSSCs), which are subspace clusters that are insensitive to the parameters of their algorithm STATPC.[8] The number of stocks in the statistical significant cluster is significantly more than expected, under the assumption that the data is uniformly distributed.

SSSCs are 2D subspace clusters $O \times A$, thus we require a postprocessing step to convert the 2D SSSCs to 3D. Given a dataset $\mathcal{D}$, which contains a set of time stamps $\mathcal{T}$, we mine SSSCs from each year $t \in \mathcal{T}$, and we try all possible combinations of them to obtain 3D SSSCs. That is, a 3D SSSCs $C = O \times A \times T$ is formed if there exists 2D SSSCs $O \times A$, $\forall t \in T$.

MIC

Correlated subspace clusters (CSCs) are insensitive to the parameters of their algorithm, MIC.[9] Unlike SSSCs, CSCs are 3D and they don't require the assumption of uniformly distributed data.

A 3D subspace cluster is a CSC when it satisfies the following criterion: the 3D subspace cluster $C = O \times A \times T$ is correlated when the values in the cluster have high co-occurrences and these co-occurrences aren't by chance.

CATSeeker

The price return of the stocks can be crucial information in clustering, but the previous three algorithms don't incorporate this information. The CATSeeker algorithm incorporates this information, and its clusters are denoted as CATSs.[10] A CATS satisfies the following criterion: the 3D subspace cluster $C = O \times A \times T$ is actionable when $\forall t \in$ T, that is, the stocks in O are similar on the set of financial ratios A; and the stocks in O have high and correlated price returns in years T.

Given a set of centroids, the optimal clusters with respect to these centroids are found. The results of the algorithm are shown to be insensitive to its parameters.[10]

On the selection of centroids, the user can set a threshold to select stocks that have good historical price returns as the centroids.

Experiments

We coded all algorithms in C++, and their codes or programs were kindly provided by their respective authors. We performed all experiments using computers with Intel Core 2 Quad 3.0-GHz CPUs with 8 Gbytes of RAM. We used Windows 7 except for experiments involving TRICLUSTER, which we performed in Ubuntu 10.10

We conducted the experiments in accordance with the parameters presented in the "Research Design" section. We set the risk-free return at $\delta_{\text{ret}} = 0$ in our experiments.

For TRICLUSTER, we fixed its minimum size parameters to $\min_O = 5$, $\min_A = 2$, and $\min_T = 3$, and varied its similarity parameters as $\varepsilon = 1$ and $\delta = 0, 0.1, 0.01$, as it's not possible to test on all possible combinations of its parameters.

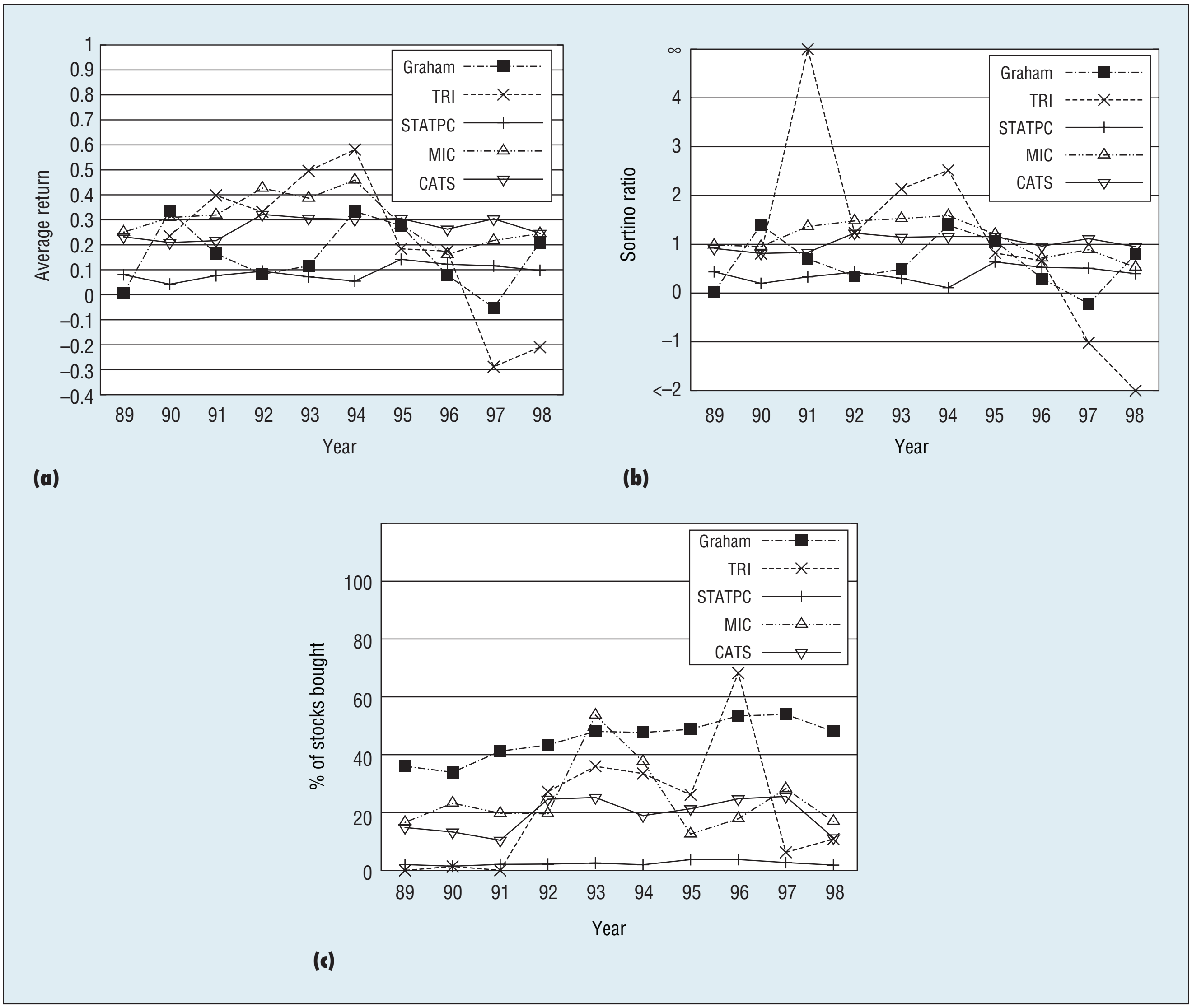


Figure 2. The different 3D-subspace-clustering strategies' (a) average returns and (b) Sortino ratios across the years. Each year on the x-axis denotes the start of a ten-year test period. (c) Percentage of stocks bought.

For STATPC, we used its default setting $\alpha_0 = 10^{-10}$, $\alpha_K = \alpha_H = 10^{-3}$. For MIC, we used its default setting p-value = 10^{-4}. For CATSeeker, we used its default setting $\tau = 0.1$; $\mu = 10$; $\delta = 0.001$; $\lambda = 0.1$; and varied $\rho = 0.2$, 0.3, 0.4, as its results are shown to be insensitive to this range of ρ.[10]

On the use of training datasets, TRICLUSTER, STATPC, and MIC used $\mathcal{D}^t_{\min_{ret}}$, as these algorithms don't consider the stocks' price returns during their clustering process. CATSeeker used $\mathcal{D}^t$, as this algorithm considers the stocks' price returns during its clustering process. For CATSeeker, a stock o is selected as a centroid if its $CAGR(o, t, t + 9)$ is at least $\min_{ret}$.

Average Returns Across Years

Figures 2a and 2b present the average returns and Sortino ratios of the different 3D subspace-clustering strategies, on testing datasets $\mathcal{D}^t$, $t \in \{1989, \ldots, 1998\}$.

TRICLUSTER-based strategy generated good positive returns in the initial seven datasets. In dataset $\mathcal{D}^{1992}$, it even has a Sortino ratio of infinity, as returns of all the stocks picked have positive returns. However, this approach generated substantial losses in the last two datasets, notably generating an 80 percent loss in $\mathcal{D}^{1998}$. Hence, TRICLUSTER produced pretty volatile results. Strategy with volatile results naturally generates high returns in the datasets when it's profitable; TRICLUSTER based strategy has the highest Sortino ratio in datasets $\mathcal{D}^{1991}$, $\mathcal{D}^{1993}$, and $\mathcal{D}^{1994}$. However, strategies with less volatile results also outperformed their volatile peer in certain years, as CATSeeker based strategy has the highest Sortino ratio in $\mathcal{D}^{1996}$,

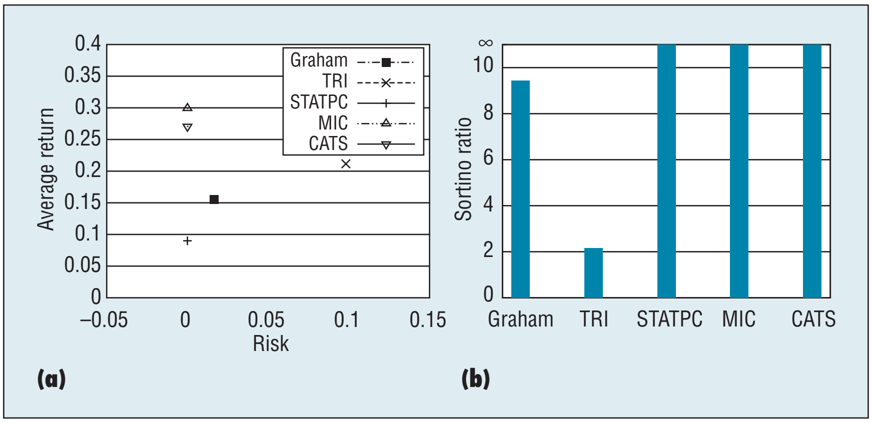


Figure 3. The results of the different strategies across the 10 testing datasets. (a) Overall average returns and risks, and (b) overall Sortino ratios.

$\mathcal{D}^{1997}$, and $\mathcal{D}^{1998}$; and MIC in $\mathcal{D}^{1989}$, $\mathcal{D}^{1996}$, and $\mathcal{D}^{1995}$.

STATPC-, MIC-, and CATSeeker-based strategies were able to generate positive average returns across the 10 datasets, which even Graham's strategy was unable to achieve (it generated losses in dataset $\mathcal{D}^{1998}$).

Percentage of Stocks Bought

Figure 2c presents the percentage of stocks bought by different strategies, based on the pool of stocks available to each strategy. The CATSeeker-based strategy bought between 10 to 20 percent of all stocks, and is better than Graham's strategy, which bought the most stocks (between 30 to 50 percent of all stocks). STATPC bought the least stocks (less than 5 percent of all stocks) but it has a lower average return than MIC and CATSeeker. TRICLUSTER and MIC are more volatile strategies, because the percentage of stocks bought by them varied across the years.

Overall Average Returns and Risks

We calculated the overall average return and risk of the different strategies across the 10 testing datasets and present the results in Figure 3a. STATPC-, MIC-, and CATSeeker-based strategies have zero risk, as they have positive average returns across the 10 testing datasets. Figure 3b presents the overall Sortino ratio across the 10 testing datasets, which shows that Graham's strategy has a high Sortino ratio. However, STATPC-, MIC-, and CATSeeker-based strategies have higher Sortino ratios than Graham's strategy. Among these 3D-subspace-clustering strategies, MIC- and CATSeeker-based strategies have higher average return and lower risk than Graham's strategy.

Summary of the Experiments

All stocks recommended by CATSeeker- and MIC-based strategies have positive returns. This is a good achievement, and CATSeeker- and MIC-based strategies generate 60 percent more returns than Graham's strategy, with no negative returns across the years (see Figure 3a). Although CATSeeker and MIC used different approaches to find 3D subspace clusters, their performances are similarly good. This suggests that good performances can be achieved by different approaches; price return is used to guide the clustering in CATSeeker, and information theory concept (correlation information) is used to guide the clustering in MIC.

We investigated the effectiveness of using 3D subspace clustering for value investing. This approach involves grouping stocks that have similarly good fundamentals (represented by their financial ratios) over the years, and then using this information to buy stocks.

We compared this approach with a highly successful value investment strategy, known as Graham's strategy. We found that two 3D subspace-clustering strategies generated

THE AUTHORS

Kelvin Sim is a scientist at the Data Analytics Department, Institute for Infocomm Research, Singapore, which is part of the Agency for Science, Technology, and Research. His research interests include financial data mining, subspace clustering, graph mining, co-clustering, and activities of daily living recognition. Sim has a PhD in computer engineering from Nanyang Technological University, Singapore. Contact him at shsim@i2r.a-star.edu.sg.

Vivekanand Gopalkrishnan is the director of research at Deloitte Analytics Institute Asia. His research interests include efficient algorithms for mining interesting item sets, subspace clustering, mining in P2P networks, outlier detection, and data warehousing. Gopalkrishnan has a PhD in computer science (data warehousing) from City University of Hong Kong. Contact him at vivek@deloitte.com.

Clifton Phua is the security and fraud analytics lead at SAS. His research interests include data mining, fraud detection, activity recognition, and intelligent monitoring. Clifton has a PhD in information technology from Monash University, Australia. He's a member of IEEE. Contact him at clifton.phua@sas.com.

Gao Cong is an assistant professor at Nanyang Technological University, Singapore. His research interests include geospatial keyword queries and mining social media. Cong has a PhD in computer science from the National University of Singapore. Contact him at gaocong@ntu.edu.sg.

60 percent more returns than Graham's strategy, with zero risk.

References

1. H.R. Oppenheimer, "A Test of Ben Graham's Stock Selection Criteria," *Financial Analyst J.*, vol. 40, no. 5, 1984, pp. 68–74.
2. B. Graham, *The Intelligent Investor: A Book of Practical Counsel*, Harper Collins Pub., 1986.
3. J.Y. Campbell and R.J. Shiller, "Valuation Ratios and the Long-Run Stock Market Outlook: An Update," *J. Portfolio Management*, vol. 24, 2001, pp. 11–26.
4. H.-P. Kriegel, P. Kröger, and A. Zimek, "Clustering High-Dimensional Data: A Survey on Subspace Clustering, Pattern-Based Clustering, and Correlation Clustering," *ACM Trans. Knowledge Discovery from Data*, vol. 3, no. 1, 2009, pp. 1–58.
5. R. Tsay, *Analysis of Financial Time Series*, Wiley-Interscience, vol. 543, 2005.
6. F. Sortino and L. Price, "Performance Measurement in a Downside Risk Framework," *The J. Investing*, vol. 3, no. 3, 1994, pp. 59–64.
7. L. Zhao and M.J. Zaki, "TRICLUSTER: An Effective Algorithm for Mining Coherent Clusters in 3D Microarray Data," *Proc. ACM Sigmod Int'l Conf. Management of Data*, 2005, pp. 694–705.
8. G. Moise and J. Sander, "Finding Non-Redundant, Statistically Significant Regions in High Dimensional Data: A Novel Approach to Projected and Subspace Clustering," *Proc. 14th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining*, 2008, pp. 533–541.
9. K. Sim, A. Aung, and G. Vivekanand, "Discovering Correlated Subspace Clusters in 3D Continuous-Valued Data," *Proc. IEEE Int'l Conf. Data Mining*, 2010, pp. 471–480.
10. K. Sim et al., "Centroid Based Actionable 3D Subspace Clustering," *IEEE Trans. Knowledge and Data Eng.*, vol. 25, no. 6, 2012.