RESEARCH

Systematic Methods for Classifying Equities

Our research explores alternative approaches for assigning sectors to the S&P 500 using stock price covariance and natural language processing.

23 AUGUST 2018
15 MINUTE READ

The Global Industry Classification Standard, or GICS, is widely accepted as the definitive approach to categorising the equity universe into industry sectors.

On September 28, the GICS Telecommunication Services sector will be renamed “Communication Services” and its remit will be broadened, increasing its size from around 2% of the S&P 500 by market capitalisation to around 10%.

At a stroke, some of the world’s largest companies will be recategorised, triggering billions of dollars of trading, as benchmarked and sector-specific products realign their holdings. This is potentially costly activity for those forced to rebalance, with the predictable trading flows prone to front-running.

Rather than being beholden to the decisions of an index provider, there are countless other ways to classify equities into sectors. We examine two systematic alternatives: one uses a covariance matrix to classify stocks based on their market behaviour; the other uses Natural Language Processing to group companies according to their own descriptions.

Boundaries redrawn

Market participants group companies into sectors for a variety of purposes, including: to invest according to thematic views (through sector-specific ETFs, for example); to construct hedges for single-stock positions; to benchmark performance; to identify closely-related companies for which fundamental indicators can be compared; and to improve estimates of risk.

As a result, GICS − the industry-standard equity classification scheme − has considerable influence on both how people view the market and how they invest. Its reorganisation on September 28 will see some of the world’s most valuable and recognisable companies switching sectors, including Alphabet, Facebook, Walt Disney and Netflix.

Such changes can spur large investment flows. The latest will see the new Communication Services sector pull in companies from Consumer Discretionary and Information Technology, both of which are tracked by popular products. The US$22 billion SPDR Technology ETF alone will have to reallocate about US$4 billion of positions, if it is to continue to track its related GICS sector – a trade driven by S&P and MSCI's reorganisation.

We have, in past research, quantified the hidden costs to passive investors due to predictable trading around stock index additions and deletions. Such slippage costs could be even higher for sector-focused products, given the scale of revisions. In the case of the S&P 500, about a fifth of the current market capitalisations of the Consumer Discretionary and Information Technology sectors is joining the new Communication Services sector.

Alternative approaches

GICS is widely accepted as the default method for classifying companies, but it is just one of many approaches, each with its own characteristics. When deciding which to use, we should consider what questions a classification is supposed to answer, and what data can best furnish those answers.

Under GICS, or the similar Industry Classification Benchmark (ICB) produced by FTSE Russell, a company's classification is decided according to analysis of its principal source of revenue and, to some extent, market perception.

Stocks in the same GICS grouping have prices that generally move together. But it can be useful to identify such stocks more precisely, particularly for hedging positions and building risk models. These applications benefit from a method that responds to market data more dynamically and systematically than GICS.

Alternatively, it may be useful to classify companies by the nature of their business: who their customers and competitors are, and to which fundamental commercial and economic changes they might respond. Under GICS, where a company can belong to only a single sector, the diversity of a modern firm may be obscured. A more flexible scheme could quantify the differences between firms and acknowledge that some sectors are more closely related than others.

In our equity trading, we are particularly interested in how investment signals apply to groups of stocks. Some signals have more statistical power when applied to a collection of stocks rather than to the entire market in all its diversity, or to a single stock with its own idiosyncrasies. Other signals may, when traded, lead to unwanted sector exposure if this is not carefully controlled.

Here we look at two alternative methods for classifying equities using data and statistics that are systematic, capable of accounting for the diversified nature of firms, change according to a consistent methodology over time and measure clearly defined properties.

One approach computes a covariance matrix to capture relationships in the movements of company share prices; the other uses natural language processing (NLP) to group companies by analysing the commentary in annual reports.

When applied to the S&P 500 (Figure 1), these two methods divide the index into sectors that include comparable fractions of the total market capitalisation and can be assigned coherent, sensible labels. Moreover, by examining the 10 largest companies in the index (Figure 2), we see they produce distinct but intuitive results.

Figure 1: Three different views of the makeup of the S&P 500

Figure 2: Alternative classifications for the 10 largest S&P 500 companies

Grouping stocks using price data

To identify stocks that might move together in the future, a natural approach is to find those that have moved together in the recent past. To do this we apply statistical clustering methods to the covariance matrix of returns.

The first step is to take price changes over some period and measure their correlation. Winton computes intraday correlations for a large collection of global assets, but here we restrict our attention to daily returns for stocks in the S&P 500 over the course of one year.

The correlation of a pair of stocks can vary between 1 (perfectly correlated) and -1 (perfectly anticorrelated). Pairs with a large positive correlation can be thought of as lying close to one another, while those with a large negative correlation are far apart. We then have a measure of distance between any pair of stocks.

Now, a group of stocks that all lie close to one another can be considered a cluster, and there are many ways to identify clusters automatically once we have a distance measure. Clusters can themselves be grouped together into larger clusters, forming a hierarchy of nested groups. Eventually, the clusters will be large enough that there will be roughly as many clusters as there are GICS sectors: we can then call these our covariance-based sectors.

In Figure 3, we show how sectors picked out this way correspond to structure that is visible in the correlation matrix. The number of sectors is arbitrary, however. Any procedure must make a choice about this number, and here we are free to choose the precise level of granularity that is appropriate for our purposes. To obtain distinct sectors of comparable size, the price data suggest using 12 groups.

We can still identify structure at a higher level, however: the tree diagram along the top of the matrix in Figure 3 shows how alike the sectors are, and how they would be grouped if we were to aggregate them into still fewer super sectors.

Since the clusters computed this way use only price data, they are objective, unlabelled and can be updated frequently. Subsequent interpretation is not essential to the method, but if we wish to gain an understanding of the patterns revealed by correlations, we need to apply our own judgment or some external information.

Figure 3: Correlation matrix of S&P 500 stocks in 2016

Divergent healthcare trends during 2016

Various structures are apparent in the covariance matrix for 2016. To take one example, healthcare, which is a single top-level GICS sector, is split into two major clusters. The pharmaceutical or biotech companies in the first group – the likes of Amgen or Gilead – behaved distinctly from the companies providing hardware and services in the second group, such as Aetna or Medtronic.

Are telecoms utilities?

The Telecommunications Services sector, meanwhile, is merged into Utilities, which seems a reasonable reflection of the underlying companies' role in the modern economy. It also provides an interesting counterpoint to the forthcoming creation of the Communication Services sector under GICS.

Grouping stocks using NLP

A covariance approach to classification is backward-looking and this is an intrinsic weakness when using sectors built from historical price data. The second data-driven method we examine is based on fundamental information. It avoids being purely backward-looking by analysing the words that companies themselves use to describe their future intentions as well as current activities in regulatory filings.

We do this using Natural Language Processing (NLP), a set of techniques that form part of the machine learning domain. This is applied to Section 1 of the 10-K annual filing required of US companies, which must include, "a description of the company's business, including its main products and services, what subsidiaries it owns, and what markets it operates in".

In analysing these filings there are two sorts of objects we want to assign to groups: first, we associate words in the documents with a limited set of dynamically learned topics, based on how often they occur together in a 10-K filing. To employ the distance metaphor again, words that often appear together in a document are considered close and are likely to be clustered together into the same topic. This means that a particular 10-K – and, therefore, a particular company – will be associated with a mixture of different topics with different weights.

At the same time, we group companies together based on the strength of their association with different topics. Companies with similar topic weights are determined to be close together and are likely to be placed in the same sector. However, the weights are useful in themselves, yielding a more nuanced view than a single sector label.

As before, we look at the S&P 500, but note that the model is trained – that is, the words are grouped into topics – using a much larger universe of companies. In common with many machine learning techniques, the more data that can be brought to bear, the better the results. We also specify the number of topics we wish to identify in advance: we choose 11, the same as the number of GICS sectors.

The resulting hierarchy of topics is shown in Figure 4. The names of the different sectors can now be assigned using the words associated with the topic. We see, for example, that there are two different healthcare topics, which are the most closely related pair of topics from the 11 chosen.

Figure 4: The hierarchy of NLP-derived clusters

This approach echoes the divergence between GICS Health Care stocks in 2016 in the covariance approach. The topic labels are generated by hand, but we see from Table 1 that they can come quite naturally from lists of the top-30 words associated with each topic. This is not always true, however: sometimes the list of words is hard to assign a single coherent label. Moreover, not all the words are informative. We see in Table 1 words such as “will”, “year” and “cost” which are either spurious or generic, which is typical for machine learning techniques.

Table 1: Top-30 words associated with each of the two healthcare topics

New media

Media companies, associated with words such as “network”, “content” and “advertising”, form a distinct sector. By subsuming the Telecommunications Services sector and including Information Technology and Consumer Discretionary companies, this new sector identifies an investment theme that cuts across different GICS sectors.

Material differences

With new sectors appearing, and the total number remaining fixed, some GICS sectors will have no counterpart in our NLP approach. Materials is one, and further analysis allows to understand why.

By restricting our attention only to companies that GICS places in its Materials sector, and selecting 30 topics, we can see Materials split into two, very distinct groups: one related to raw materials, the other to processed materials. NLP does not support merging them together into a single topic. Rather, “Materials” companies are associated variously with retail, hardware and energy topics.

What is Amazon?

Amazon is an interesting test case for classification schemes: not only is it a large and diverse company that is closely watched by investors, but its business – and the market’s perception of its business – has changed over time.

The covariance perspective

In Figure 5 we show the correlation in 2016 between Amazon and the other stocks in the S&P 500, colour-coded by GICS sector. We also highlight, with crosses, all the stocks included in the same covariance-based sector as Amazon. This sector includes both Consumer Discretionary and Information Technology stocks.

If we examine more closely the stocks with which Amazon is grouped, we find other internet retailers, such as Priceline, along with a scattering of brands: Nike, Starbucks and Under Armour. Its Information Technology companions include Facebook, Alphabet, Netflix, Microsoft and other software companies,

but not Apple. This highlights the difficulties we can face in trying to fit a company into a predefined list of sectors.

Figure 5: Correlation of Amazon with the rest of the S&P 500 during 2016

The NLP perspective

When using 10-K filings, we have new information on a company's sector membership once per year. Updates therefore happen more dynamically and more systematically than changes in GICS classification, which has historically reacted only seldom to developments in the market environment, despite annual reviews.

Figure 6 shows the weights given to the different NLP topics in Amazon's 10-K filings between 2007 and 2016. It starts this period describing itself predominantly as an IT company, associated with the software and hardware topics, with less emphasis on retail.

In recent years, however, it has clearly described itself as being in the retail business, with IT as a secondary topic, and media making a small but increasing contribution. Reassuringly, there is negligible weight given to categories that our judgment tells us are peripheral to Amazon, such as topics associated with finance and healthcare.

Figure 6: Evolution in the topics Amazon uses to describe its business

Amazon was in the GICS Consumer Discretionary sector over the whole period. This looks fair, if we are forced to choose a single sector. It obscures the fact, however, that many of its customers buy consumer staples or cloud computing services from Amazon. Moreover, many investors still think of it as being a tech company. The covariance of its price changes and the contents of its regulatory filings both reflect this more faithfully.

No default choice

In moving away from a somewhat imposed, discretionary, one-size-fits-all scheme for stock market sector classifications, it is possible to adopt different approaches for different applications.

We have shown the benefits of flexible and data-driven schemes of classification, but caution is required. Noise in the data may throw up potentially spurious relationships between stocks or identify counterintuitive groupings that are unlikely to persist.

Furthermore, if we simulate a trading system that makes use of proprietary, dynamic sectors, we must decide how we would have defined them in the past and how they would have changed through time, introducing the possibility of hindsight bias. A long history may be difficult to reconstruct, given changes in the data available to feed into machine learning algorithms, though we note that other sector definitions, including GICS, may also have relatively short histories.

None of the methods we have discussed should therefore be considered a default choice, as the GICS often appears to be. Rather, when deciding how to classify a company, we ought to start by considering what question the classification is meant to answer. Then we can assess what data is required to make the decision, and which statistical methods are most appropriate.

Download the PDF

SYSTEMATIC INVESTING EQUITY INVESTING

Related Articles

The Role of Sector and Country in Stock Returns

Does globalisation mean that stock returns are related more to sector than country performance?

Seasonal Volatility and the Multiplicity Effect

October has been the most volatile month for stocks on average over the past 87 years. Is this due to chance?

The Hidden Costs of Global Index Tracking

We find hidden costs of about 10 basis points per year for the naïve global index tracker.

Winton is a global investment management and data science company. Founded in 1997, Winton's business is grounded in the belief that the scientific method can be profitably applied to the field of investing.

Follow us

Subscribe

Subscribe to the latest updates from Winton.

Your email