Financial ratios and stock returns reappraised through a Topological Data Analysis lens

ABSTRACT

Firm financials are well established predictors of stock returns, being the basis for both the traditional econometric, and growing Machine Learning, asset pricing literature. Employing topological data analysis ball mapper (TDABM), we revisit the association between seven of the most commonly studied financial ratios and stock returns. Upon outlining the methodology to the finance literature, this paper offers three key contributions to the study of asset pricing. Firstly, the characteristic space is visualised to showcase non-monotonic relationships in multiple dimensions that were as yet unseen. Secondly, the means through which neural networks and random forest regressions fit stock returns is also visualised, showing where Machine Learning is contributing to understanding. Finally, an initial application of TDABM for the segmentation of the cross-section is posited, with significant abnormal returns identified. Collectively these three expositions signpost the value of TDABM for financial researchers and practitioners alike. Scope for benefit is limited only by the availability of information to the analyst.

1. Introduction

Stock returns are intrinsically linked to firm financial ratios, with a vast literature seeking the optimal model of exactly how. Driven by data abundance, computational power, an ever expanding set of data science techniques, and an appreciation of methods for data reduction, mining for mispricing continues to accelerate. From the data science perspective, it has been the Machine Learning family employed in considering the “factor zoo” (Stambaugh and Yuan 2017; Feng et al. 2020) of potentially relevant characteristics that has dominated (Green et al. 2017; Kozak et al. 2019; Gu et al. 2020). Critics contend Machine Learning (ML) is too much “black box” about the exact mechanism linking input characteristics to output returns, meaning adoption of ML is therefore not as rife as the fit improvements ML offers suggest it should be (Arnott et al. 2019). However, data science may also unearth new open perspectives on the whole process of determination of stock returns. This papers delivers new insight through the application of Topological Data Analysis (TDA), an approach widely adopted in the physical sciences, which has much to offer to the study of stock returns. Specifically it is demonstrated that mapping firm characteristic space through TDA Ball Mapper (TDABM) yields clarity on both the stock return to firm financial ratio relationship and the ways in which ML is fitting thereto. TDABM can offer a tool to leverage the non-linearity for investment. The results of this paper are supported by seven appendices, which are available from the corresponding author on request.

TDA, after the seminal work of Carlsson (2009), views data as a point cloud, a multi-dimensional space in which each observation has its co-ordinates. Inference from that point cloud follows the logic of established inference from two dimensional clouds, scatter plots. Benefits of seeing shape are established in the classic Anscombe's quartet (Anscombe 1973) that prove regression lines, correlations and summary statistics alone are not sufficient for understanding the actual relationships within data[1]. TDABM offers a means through which to visualise the multidimensional point cloud and hence to abstract information therefrom. By covering the cloud in equal radius balls, TDABM graphs give measures of connectivity, density and outcome distributions. This paper has a point cloud from firm characteristics and observations which represent stocks; a ball is then a portfolio of neighbours in the characteristic space. As well as visualising the cross-section, TDABM has a natural interpretation upon which we highlight an agenda for asset pricing research and practice.

Contributions of this paper are threefold. Firstly this is the first paper to apply TDABM in the study of stock markets, a major element of the paper then being the demonstration of the technique in a finance context. Secondly, further evidence of the effect of monotonicity within the firm characteristic space is provided, adding visualisation to the evidence base. Our results demonstrate how non-linearity and mispricing from linear models should not be conflated. Finally, we show how segmentation of firm characteristic space with TDABM offers potentially investable strategies that are mispriced by the market. Primarily focus is on the ability of TDABM to provide new perspectives on the cross section of stock returns, directing the research agenda in

[1] A contemporary version using multiple graphs with the same regression lines, correlations, means and standard deviations can be found in the datasaurus of Matejka and Fitzmaurice (2017).

complementarity with ML.

Fama and French (2020) identifies models premised on the characteristics of firms as being on the cross-section, recognising the centrality of Fama and MacBeth (1973)[2]. Almost exclusively it is defined that an anomaly exists where it is possible to generate abnormal returns by taking a zero cost trading position, longing one portfolio and shorting another. In such cases the portfolios are assumed to sit at the two ends of the distribution of the characteristic; a methodology developed in Fama and French (1992) and Fama and French (1993). Fama and French (1993), Ou and Penman (1989), Basu (1977), Haugen et al. (1996) and Fama and French (2015) all fall into this mould of sorted portfolios and inspire the selection of common anomalies used as illustrators in this paper. There is a plethora of further works in this vein which have identified anomalies from firm characteristics, each premised on the idea of segregating stocks under the implicit monotonicity assumption that the highest and lowest returns would be at the top and bottom of the characteristic distribution. Works in this vein include Hou et al. (2015), Stambaugh and Yuan (2017) and Daniel et al. (2020). Whilst identification tests for new anomalies are suggested to be tightened (Harvey et al. 2016), the basic protocol for identification remains the same (Pukthuanthong et al. 2019). Using TDABM we demonstrate that the requisite montonicity is limited, but that the information within stock characteristic space maintains potential to segment the cross-section.

Patton and Timmermann (2010) and Romano and Wolf (2013) provide early examples of tests for mispricing that do not require the implicit assumption of monotonicity. Response to the modelling implications of non-monotonicity is made by Shively et al. (2009) and reviewed in the contemporary context by Fisher et al. (2020). However it is in the empirical application of Machine Learning models where the non-linearity has been most directly embedded. Within works like Kozak et al. (2019), and Gu et al. (2020) the process begins with a set of characteristics that have been identified as potential anomalies before a data reduction process creates latent factors for the second phase of analysis. Principal components analysis has also been adopted for dimension-

[2] Fama and French (2020) then refers to those studies explaining returns using factor models like Fama and French (1993) or Fama and French (2015) as being time series.

ality reduction and again creates a non-linear relationship (Lettau and Pelger 2020). The reduced set may then be fitted by both linear and non-linear models. Through TDABM we show how linear models have failed to explain the cross-section, but that equally fitting neural network and random forest regressions on the reduced set of factors produces very similar residuals. A linearity within the data is suggested; the residuals pattern stems from factors omitted from our limited set. TDABM graphs, and measures thereon developed in this paper, deliver value in demonstrating consistency between approaches across the space.

Fama and French (1993) and subsequent works premise on the segmentation of assets into portfolios by quantile breakpoint. However as the number of dimensions increases so the ability to segment on multiple firm characteristics diminishes. Bryzgalova et al. (2020) proposes a mechanism through which a random forest regression may be used to determine the portfolio cuts, an approach which offers promise. Likewise, there is preliminary work on the creation of peer groups of stocks using k-means clustering (Hartigan and Wong 1979). Both Snow (2020) and Ge et al. (2021) that suggests opportunities for arbitrage. Potential for k-means to be used to select stocks was first identified in Liao et al. (2008) but little had been explored in the literature. In the final part of this paper TDABM is explored as an alternative means of segmentation of data, with alpha demonstrated to result. Unlike k-means and Random Forest which produce irregular clusters, TDABM has the advantage of using common radii. Helping avoid accusation of data mining it may be noted that balls are model free, just as are the equally spaced quantiles used historically in the finance literature[3]. Irrespective of the cluster shape, or portfolio formation process, monotonicity is not required in any approach, rather the quest is to identify parts of the space that are consistently offering high and low returns and then arbitraging across.

The remainder of the paper is organised as follows. Before outlining the method, data for this study is introduced in Section 2. A theoretic and context relevant expo-

[3] Where New York Stock Exchange (NYSE) quantiles are used there are more stocks in the lower size quintiles owing to the larger size of stocks listed on the NYSE. However, the choice of quantiles in the literature is always equally spaced. Fama and French (1993) and others use the top 30% and bottom 30% for their estimation whilst Stambaugh and Yuan (2017) uses the top 20% and bottom 20%. In each case the split is symmetric. To this end we adopt quantiles as simpler terminology than referencing the NYSE each time.

sition of the TDABM algorithm is provided in Section 3. Section 4 highlights insights from the characteristic space. Contrasting the fit of linear and ML models, Section 5 provides the second contribution of the paper. Exploring the mispricing of balls from the TDABM diagrams, Section 6 evaluates the lessons for theory and practice in asset pricing. Finally, Section 7 concludes.

2. Data

Taking stock data from the Center for Research in Stock Prices (CRSP) and Compustat, firms' financial ratios are constructed using code from Green et al. (2017). This paper then uses size and book-to-market ratio (Fama and French 1992, 1993), profitability and investment (Fama and French 2015), earnings-to-price (Basu 1977, 1983), the dividend yield (Black and Scholes 1974; Fama and French 1988), and cashflow-to-price (Haugen et al. 1996). Details of the construction of the variables are provided in the notes to Table 2[4]. This paper considers data relating to stock returns between January 1978 and December 2018, but only a section of this is used in the illustrations. In the examples that follow five year-month combinations are used, being June 1978, June 1988, June 1998, June 2008 and June 2018. These specific months are motivated as ten year intervals either side of the turning point in US GDP in the global financial crisis. In what follows, these ten year intervals are used for illustration but inferences are robust to the choice of other months. For the portfolio analysis we use the full sample. All characteristics are winzorised at the 0.5% level. Table 1 provides summary statistics for the 1980 to 2018 sample, as well as for June 2018.
—- Table 1 Here —-

For all variables summarised in Table 1, the interquartile range is small relative to the overall range. It is later shown how this creates a common mass at the centre of the Ball Mapper plots. Because of the variation in the levels of the seven common anomalies, in common with applications of PCA and ML, normalisation is applied to place each on the scale $[0, 1]$ prior to running the Ball Mapper analysis.

[4] A further discussion of variable construction is available in Appendix A on request.

3. TDA Ball Mapper

To motivate the applications that follow, an overview of the TDABM algorithm is given. To then nest TDABM within the financial data science literature a brief review of the similarities, and critical differences, between TDA, TDABM and ML is provided[5].

3.1. *Ball Mapper Algorithm*

Understanding high dimensional and complicated data is a keen pursuit for the Machine Learning and Data Science community. The input to the presented analysis is a d dimensional point cloud X, i.e. any given datapoint is defined by its coordinates on the d axes $(x_1, x_2, ..., x_d)$. In addition, for every pair of points $x, y \in X$ the metric or a symmetric similarity measure $dist(x, y)$ is given.

The main purpose of the TDABM algorithm is to create a cover of the space X with a collection of potentially overlapping sets $C_1, \dots, C_n$ such that each C_i is a subset of X[6]. In TDABM these sets are balls constructed around a chosen collection of landmark points. The radius is the only parameter of the algorithm. The number of sets, n, is determined as an outcome of the algorithm as described below. Moreover it is then implicit that the points gathered in each C_i are geometrically close.

Given the collection $C_1, \dots, C_n$ that cover X, the vertices of the TDABM graph G correspond to the elements of the cover, i.e. each vertex i of G corresponds to C_i. The vertices i, j in G are joined by an edge if and only if $C_i \cap C_j \neq \emptyset$, i.e. the corresponding elements of the cover have a non-empty intersection. One can define *weights* of both vertices and edges of G. In this paper, for simplicity, we will use uniform weights for edges. The weights vertices of G however will represent the density of the point cloud X. To achieve that purpose, the size of a vertex i will be proportional to the cardinally of C_i.

[5] For interested readers an intuitive guide to construction of a TDABM plot for two variables is available as Appendix B on request. An introduction to the use of the BallMapper R package (Dlotko 2019) may also be obtained on request from the authors as Appendix C.

[6] TDABM is inspired by the *Reeb* and *Conventional Mapper graphs as well as nerve of a cover.* Interested readers may also see Carriere and Oudot (2018) for a further review of the original mapper software.

It remains to explain how the cover $C_1, \ldots, C_n$ is obtained. There are a number of ways to get it and in what follows focus is placed on the most intuitive one. This method for obtaining a cover is based on so called *ϵ-net* of the space X. Consider a collection of points X equipped with a distance or similarity measure $dist : X \times X \to \mathbb{R}$. For a given $\epsilon > 0$, an ϵ-net is a subset of points $X' \subset X$ such that for every $x \in X$ there exist $x' \in X'$ satisfying $dist(x, x') \leq \epsilon$. Suppose that $x_1, \ldots, x_n = X'$. Then we can set $C_i = B(x_i, \epsilon)$ for every $i \in \{1, \ldots, n\}$ to obtain a cover of X.

$\epsilon-$net is not unique and in this work it is chosen using the following randomized procedure. The first point x_1 of X' is taken at random from X. Subsequently all the points in X that are not farther away than ϵ from x_1 are marked as *covered.* If there is still a point $x_2 \in X$ that is not covered, it is added to X' and the procedure is repeated until all points in X are covered. Note that some of the points in X may be covered by balls centred at multiple points in X' – they will give rise to edges of G. The number n of cover elements depends on ϵ as well as the process of selecting points in X'. No closed expression that determines n given X and ϵ is known, therefore the number n should be treated as an output of the algorithm computing ϵ net.

Note that, in the construction above, all points that come from a ball of radius ϵ are at most 2ϵ away from each other, and therefore can be considered as being *geometrically close.* The name of the TDABM algorithm is accredited to the fact that cover of the space X used herein is made by using balls (in a given metric or similarity measure $dist$). The radius ϵ, point cloud X and the distance $dist$ are the only parameters required by the TDABM algorithm. One should think about ϵ as a level of resolution, with respect to $dist$, that one wants to use to examine data; small values will show detailed structures in X, while large ones will give the overall idea of the layout of X.

The presented construction gives a way of representing the shape of a complicated and high dimensional point cloud X by using an abstract graph G. In addition thereto it often happens that the points of X are accompanied by a function $f : X \to \mathbb{R}$ that is of interest. Given such a function f let us define the function $\hat{f}$ induced by f on the TDABM graph G; $\hat{f} : G \to \mathbb{R}$. For every cover element C_i, the value $\hat{f}(i)$ is an average value of f on elements of C_i. Note that for smooth f's, its value restricted to C_i will not vary much; f will be close to $\hat{f}$.

The function $\hat{f}$ on the TDABM graph is subsequently visualized by an appropriate colouration of the vertices of G. Doing so commends the TDABM algorithm as an advanced way of plotting relations $(x, f(x))$, for which x is sampled from a complicated and high dimensional set X and f is a scalar valued function defined on X. This tool often allows to locate and understand various complicated relations between variables in X. This deeper appreciation of the characteristic to outcome mapping is one of the most important motivations for the empirical examples that follow.

By observing variation in outcome $\hat{f}$ across the space, represented by G, interesting cases may identify themselves. For example restricted areas where f varies between its low and high values are obviously of interest as they identify combinations of the axis variables and recognise the joint effect of small variations amongst them. This is seen in context in the applications. Interpretation is thus summarised as first an understanding of the shape, second an exploration of the contribution of the axes of X, and third an analysis of any seeming statistical anomalies that large variation in f would represent within compact regions of X.

Balls' knowledge of their constituent data points permits further analysis of variations within, and between, balls. Whilst the interpretation of within and between variation does not serve as a radius selection criteria, as it would for the number of clusters in k-means, measures of this variation can still be used for analysis. Connectivity and size variations inform on the joint density of the distribution of characteristics; the fixed nature of a ball's radius directly equates more points with greater density. Existence of overlap permits combinations of balls to represent density over a wider space.

3.2. *Ball Mapper and Machine Learning*

TDA is the study of the shape of data, understanding the complete dimensionality of the dataset at hand. Model free metrics are produced, in exactly the same way that statistics produces metrics of data such as correlation and variance. TDA then speaks to financial data science as a means "for exploring and critically assessing new datasets to explain behaviour" (Brooks et al. 2019). Born of mathematics, TDA also speaks to Brooks et al. (2019)'s suggestion that financial data science should be"an inter-disciplinary process which is rigorously and repeatedly exploring and explaining the variance in all relevant data sets to enhance financial decision making" (Brooks et al. 2019, p1629). TDABM is one part of the TDA toolkit which has been developed after Dłotko (2019) for the precise purpose of exploring and understanding datasets.

In the terminology of Hoepner et al. (2021), TDA and TDABM are white box approaches. Everything that may be inferred from the visualisation of the data, or measures on the TDABM graph, are fully traceable back to the individual data points that comprise the underlying data set. However, the aim is not to demonstrate causality and so, as we demonstrate in this paper, TDABM is more suitable as a lens to show how other models, including ML are showing causality. Where ML models take a set of inputs and develop a mapping to a stated outcome, TDABM offers a means to visualise how that outcome varies across the full input space. In this way it is captured which combinations of input characteristics are associated with different outcome levels. This paper also shows how TDABM may then show the differences between ML predictions and true stock returns vary across the firm characteristic space. The black box nature of ML means we may not know exactly why regions of the data set have poor fit, but we may be guided to consider those regions in the evaluation of ML models. TDABM adds to the explainability and augments discussions of significance and relevance in both the statistical and economic/societal sense (Hoepner et al. 2021).

TDABM is primarily a means of creating a topologically faithful representation of a multi-dimensional dataset which may be projected onto a two-dimensional space for analysis. Here we may consider TDABM as performing a dimension reduction which preserves all aspects of the distributions of the underlying data set. Contrasting with dimension reduction algorithms, such as principal components analysis, we see TDABM preserves explainability precisely because information is not lost in the reduction. Like neural networks the output of the TDABM algorithm is dependent upon a seed. Unlike neural networks, multiple repetitions can be easily performed to get confidence intervals around any metric with little computational cost. Because the underlying data is not impacted by the seed choice, TDABM results are very stable, especially in large datasets.

The relationship between TDABM and ML may therefore be seen as one of complementarity. For the financial data scientist TDABM becomes a useful tool to augment understanding of data sets. As we demonstrate in this paper, there are many add on benefits to using TDABM, such as the ability to create investable portfolios, which exploit the explainable AI nature of TDABM.

4. Visualising Returns

First focus is on the ability of TDABM to visualise stock returns across the multidimensional set of firm characteristics. Here a two variable case after Fama and French (1993) is followed by the full seven axis set of common anomalies applied throughout this paper. In what follows we work with one representation of the TDABM algorithm at a single radius. Results are highly robust to landmark selection and are qualitatively similar at other radii[7].

4.1. *Bivariate Example: Size and Book-to-Market*

Fama and French (1993) three factor model (FF3) proposes that the CAPM be augmented with sorts on size and book-to-market. These latter two variables are then the firm characteristics used in this section. First visualise this two-dimensional point cloud as a scatterplot, colouring the points according to which quintile of the return distribution they belong. So doing allows the first formation of inference about the joint relationships between size, book-to-market ratio and returns. Figure 1 shows the resulting cloud.
—- Figure 1 Here —-

As Figure 1 demonstrates, the range of book-to-market ratios is highest amongst the smaller stocks. Five colours make it easy to see that the smaller stocks are delivering

[7] A discussion of the robustness is available as Appendix D on request from the authors.

lower returns, but that there are also a higher number of black dots in the lower size range. To the right sit more reds and greens, being stocks whose return is around the median and upward. Albeit noisy, discerning any pattern amongst this data is challenging. From a shape perspective the central mass is clear, as is the impact of winzorisation.

Each TDABM representation of the set is clearly abstract from the scatter plot of the same two variables. To see this consider again June 2018, colouration of the TDABM plot in panel (a) of Figure 2 being the return on the stocks in July 2018. For this section we use a radius of $\epsilon = 0.1$ to trade off the noise witnessed in Figure 1 with the potential loss of detail that comes from larger balls. For visualisation 0.1 works well and there is an intuitive link to the 5 by 5 bivariate sorts used in Fama and French (1993) since holding one variable constant the maximum difference between two points in the same ball is 0.2, or 20% of the overall range of the axis.

Correspondence of the shape with the cloud in Figure 1 can be seen, but by colouring by the average value of the two axes panels (b) and (c) confirm that the left of the TDABM plot is where the largest firms are and the range of BM values is lowest. To the right of the TDABM plot are the smaller firms, with the higher BM values being those in the arm to the top of the shape. Amongst the smallest firms we see high average returns where BM is high, but low where BM is low. Such concords fully with the value factor of Fama and French (1993). Where the BM values are low along the bottom of the graph the larger sized firms offer positive returns, a pattern which is repeated right up the shape. Theory suggests that it should be smaller firms offering the highest returns, but in this joint space it is only those small firms with high BM that do so. In the plot there is a broad monotonicity from small-low BM firms in the bottom right, to large or high-BM firms to the left and top. Figure 4.1 shows the same three panels for June 1978, June 1988, June 1998 and June 2008 where the neat monotonicity is not present.

—- Figure 2 Here —-

—- Figure 3 Here —-

In 2008 the highest returns may be found in the centre of the mass, with similar returns appearing on the left and right extremes. However, the presence of the very large negative return in ball 41 means that the colouration is less revealing. Common practice in the cross-section is to winzorise the firm characteristics but not the returns, therefore TDABM here is reminding us of the limited variation available then for any model fitting July 2008 returns on June 2008 characteristics. Fama and MacBeth (1973) regressions are based on average coefficients from monthly fits and so such concerns as visualised here are important. 1998 sees more variation, the highest returns coming in the small firms to the bottom left and the lowest returns coming in the larger mass to the middle right that have median size and low BM. In the 1998 data high BM is to the left of the plot but returns are not larger in this space. Tracking along the right side, where size moves from small to large the returns first fall then rise, this is a "U" shaped relationship that certainly violates monotonicity. 1988, panels (g), (i) and (j), present a very mixed picture with returns seemingly more randomly distributed over the space. Meanwhile 1978, panels (h), (k) and (l), have a randomisation of returns amongst the smaller stocks on the right, whilst the centre and left, mid to large stocks with low BM ratios, have returns around the median.

Opposite the scatter plot of Figure 1 the TDABM plot is easier to read, but the use of a radius of 0.1 has reduced the detail. That such different patterns appear in the five Junes considered is also a worthy reminder that the theoretical monotonicities are not always present. Highest returns have not always been found in the same part of the space. When considering TDABM and investment subsequently this constant switching of the part of the space in which the highest returns lie will be critical to the ability of TDABM to generate alpha. Recognising the high standard deviations of returns on the time series factors shown by Fama and French (2020) it is unsurprising that we see this; the contrast here is that we are seeing the switching in multiple dimensions.

4.2. *Common Anomalies*

TDABM exerts its advantage when there are more dimensions than can be visualised by scatter plots alone. Seven common anomalies discussed in Section 2 present such a situation. Panel (a) of Figure 4 shows how the extremes of returns do indeed occur in outliers, but that there are high returns to be found on many arms of the main shape. Understanding of the characteristics to which these arms speak necessitates a look at the colouration by axis plots of panels (b) to (h). What may be immediately taken from the colouration of panel (a) is that there is not a monotonic distribution of return across the joint distribution of these variables.
—- Figure 4 Here —-

High returns may be found above the large mass in the centre of the plot. Working through panels (b) to (h) reveals these are small firms with high BM, mid-range profit, investment and cashflow, low dividend yield and mixed earnings. Closer inspection shows the variation in returns matching to the variation in earnings to price ratio here. To the lower right there are also some high returning balls attached to the main shape. Here the size is around the median, but there is some link between size and returns. BM is likewise mixed but with smaller variation and hence less alignment to panel (a). Of note is the red ball 82 which delivers strong negative returns in the subsequent period and aligns with small size, low BM, low profit and low earnings-to-price. Variation through these balls is in spite of the high dividend yields offered by all. As the data lies behind these plots we may query the database to find out what stocks are driving the pattern; consideration of events that occurred to that firm in June 2018 may then be made.

Plots of June 1978, June 1988, June 1998 and June 2008 also show this value of seeing combinations of characteristics. As in the bivariate case the part of the space that is offering the best returns changes[8]. This is consistent again with the message from the data in Fama and French (2020) and other studies on the time series of risk factors. Here we are simply showing the changes simultaneously in a multidimensional space for the first time.

[8] These four plots and associated discussion are available on request from the authors as Appendix E.

4.3. *Ball Mapper for Visualisation*

Progressing through examples we see how visualising stock returns across the joint distribution of firm characteristics produces additional information. However, the inconsistency between periods means that necessarily these illustrations are just a first step into appreciating the full picture. TDABM offers opportunities to take measures on the space and to visualise more than just the observed returns. It would not be expected that an investor would sift through historic plots, rather the TDABM plot may be of interest for the most recent period. We next consider applications where the metrics on the TDABM plots can be used without reference to the visualisations themselves.

5. Regression Comparison

TDABM is primarily a visualisation tool, albeit it being one from which we can take metrics on our data. Moving beyond the presentation of returns, this section showcases TDABM as a way to understand residuals. In a simple ordinary least squares (OLS) exercise where the aim is to fit a linear model a la Fama and MacBeth (1973) for the excess returns one month ahead of the characteristics, we identify potential non-linearity of returns with respect to characteristics. Secondly, using a training dataset comprising the previous 11 months, it is shown that the non-linear models of Machine Learning may improve model fit across the characteristic space. Hereby, TDABM as a tool to understand the performance of Machine Learning models is introduced.

5.1. *OLS Regression*

In understanding the relationships between firm characteristics and stock returns it is typical to employ OLS regressions. TDABM can contribute to the interpretation of the messages therefrom, informing on model fit and suggesting areas where better specification could be achieved. To illustrate this argument the following model is

—- Table 2 Here —-

estimated:

$$R_{it} = \alpha + \beta_1 \; Size_{i,t-1} + \beta_2 \; BM_{i,t-1} + \beta_3 \; ROE_{i,t-1} + \beta_4 \; Invest_{i,t-1} \\ + \beta_5 \; EP_{i,t-1} + \beta_6 \; Cash_{i,t-1} + \beta_7 \; DY_{i,t-1} + \omega_{it} \qquad (1)$$

Data is taken from the time period to which TDABM is applied, and is considered as $t-1$ in the fitting of returns from time t[9]. Variables are as defined in the data section and ω_i is the Newey-West standard error fitted with 6 lags (Newey and West 1987). Estimating equation (1) for the five Junes discussed thus far produces Table 2, where the figures in parentheses are the absolute values of the Newey and West (1987) robust t-statistics for a test that each coefficient is equal to zero. Returns are expressed as percentages.

Residuals from these regressions may be readily used as a colouring variable for TDABM, as seen in Figure 5. Outliers in the parameter space have some of the highest residuals, but it is in the connected components that the interest lies. Because there is value in both appreciating the direction of prediction errors and their absolute magnitude, two TDABM plots are provided for the 2018 data. These errors are again those generated following Newey and West (1987).
— Figure 5 Here —

In Section 4 it was discussed how the lower right arm of the shape produced very varied returns despite the close proximity of the balls within the arm. This is then picked up in the directions of the residuals in panel (a) of Figure 5 and in the magnitude of the absolute residuals in panel (b). An interesting observation of panel (b) is that the lowest absolute residuals are not in the main mass of the data, rather they are in the left and the outliers. Herein lies a suggestion that the characteristics have different importance in different parts of the plot. ML models have the ability to deal with such situations, especially random forest (RF) regressions (Breiman et al. 1984). Plots of,

[9] In previous sections we refer to the time period of the plot and the returns used as colouration coming from the next period.

and related discussions about, the regressions for 1978, 1988, 1998 and 2008 deliver similar conclusions on the variation of fit within the characteristic space[10].

Identifying areas of poor fit within the characteristic space has a further advantage in that the analyst can consider which would be the optimum interaction terms to introduce to the regression model. Given that there would be significant correlation within the full set of interaction terms, being able to identify the best without relying on regression based optimisation techniques circumvents many of the fitting problems the multicolinearity would otherwise have created. Here discussion has already identified the relevance of interactions between earnings, profit and dividend yield in that lower right arm where residuals are high. These are though small balls and aiming at the terms to improve fit therein may damage fit elsewhere. Depending on the goal of the analyst, targeting better fit in the larger balls may be desirable; such could be obtained by greater winsorization. Extent of winzorisation is a further trade off between accuracy and desire to fit the majority of stocks.

5.2. *Out-of-Sample Regression Fit*

Taking training data from the previous 11 months prior to the month being analysed, let us now explore how Machine Learning methodologies may offer improved fit, visualising the resulting residuals using TDABM. Our model is of the form:

$$r_{t+1} = \alpha + \theta(f_t) + \upsilon \tag{2}$$

Where f_t are the firm specific characteristics observed at time t and r_{t+1} are the excess returns in the subsequent period. An OLS model for (2) allows each characteristic to enter independently in a linear form. However, as highlighted by our TDABM analysis, and studied in the literature post Patton and Timmermann (2010) the precise form of $\theta(f_t)$ need not be either linear, or monotonic. Prominent in the modelling techniques to respond to non-linearity and non-monotonicity are RF regressions after Breiman et al.

[10] These plots, as well as the associated discussion of model fit, are omitted for brevity. The relevant material is available on request as Appendix F.

—- Figure 6 Here —-

—- Figure 7 Here —-

(1984) and neural networks (NN) as employed by Gu et al. (2020). Herein we do not attempt to identify the optimal set of characteristics, or to identify the best functional form for equation (2), rather we seek to demonstrate how TDABM can visualise the gains made by Machine Learning models in different regions of the domain.

Figure 6 presents a TDABM visualisation of the absolute residuals from three models. Panel (a) features a NN where the architecture is either one hidden layer of 4 neurons, one hidden layer of 3 neurons or two hidden layers with 4 and 2 neurons. As in Gu et al. (2020) selection of these architectures is based on the geometric pyramid rule of Masters (1993). In practice we see little deviation in fit between different architectures, but proceed with whichever of the three produces the lowest mean squared prediction error for the training data. Below this, panel (b) provides the OLS residuals for parameters fitted on the training set. Finally, panel (c) is formed from a RF regression with 1000 trees.

Contrasts between the models are far smaller than might be expected given the fitting advantages of ML models to non-linear relationships. As in the single month in-sample regressions of the previous section, the regression models are not fitting as well as some of the extremes of the plot. Now we are seeing that the ML approaches are not delivering significant improvement for explaining next month stock returns. On a ball-by-ball basis ML models produce higher residuals, but the absolute residuals of the NN do constitute improvement over the OLS and RF approaches[11]. Summarising, Table 3 compares the root mean squared error (RMSE) of the three models at 5 year intervals, highlighting similarly that the ML do not always give improvement out of sample. We also report the average for the 41 years between 1978 and 2018 inclusive. Figure 7 further emphasises the point, showing how the RMSE values track each other throughout. Only in 2015 does the RF RMSE drop below the OLS. Emphasis is made a this point that the models fitted were for illustration with default settings rather

[11] A discussion of the pairwise t-tests on the colouration of each ball is included in Appendix F, which is available on request from the authors.

—- Table 3 Here —-

than being exhaustive searches of the parameter space. On no occasion does the RF obtain the lowest RMSE, and the differences between NN and OLS are small. Through the TDABM it has been shown that it is within a limited part of the space where this lack of fit has occurred.

5.3. *Summary*

Where returns are non-monotonic across characteristic space we would be unsurprised to see patterns in the residuals of models that presume monotonicity. Our second contribution has been to highlight that the non-linear models, NN and RF, are presenting very similar models to the standard OLS. With both ML algorithms optimising their use of inputs their returning close to linear residuals suggests that they too are positing a linear relationship between characteristics and returns. There is scope to look at more time periods, larger training samples, consideration of different architectures and the inclusion of more characteristics. Evaluating the lessons that can be learned on model fit from TDABM offers a fruitful avenue for further research. From a financial perspective, recognising the continued value of OLS is of use, as will be the efforts to find models that can beat the linear, even if only in subsets of the characteristic space.

6. Investment and Ball Mapper

Segmentation of the stock characteristic space into balls begs natural comparison to the sorted portfolios commonly applied in empirical asset pricing. Treating each ball as a portfolio we ask whether there are significant differences in the returns of the balls within a given TDABM diagram[12]. Traditional application of NYSE quantile breakpoints, as used to construct sorted portfolios in Fama and French (1993), is premised on the belief that segmenting the characteristic space may yield beneficial investment insight. Where sorted portfolios focus on either univariate or bivariate sorts, TDABM is able to segment the cloud on all dimensions simultaneously. Herein

[12] We thank an anonymous referee for the suggestion to try mispricing across balls.

we explore the investment insights and implications from such a segmentation.

Formally we conduct two analyses. Firstly, using the five example years we explore variation in the characteristics that produce the ball with the highest (lowest) return in subsequent periods. Secondly, taking the full sample we identify the portfolios that produce the biggest returns in sample and ask whether selecting such stocks can potentially deliver alpha. An exploratory analysis, the approach relies heavily on the presence of momentum in the markets to ensure that returns persist from the time period in which the BM graph is constructed to the time period in which the returns are evaluated[13].

Analysis of investment potential makes use of the full sample from 1 January 1978 to 31 December 2018. In the case of annual rebalancing only the complete years are used reducing the sample slightly to 1 July 1978 to 30 June 2018.

6.1. *Bivariate Examples*

A four step process is employed. Firstly, the TDABM cover is generated at a given radius from the relevant anomaly data at time t. Those balls with 30 or more stocks are considered as potential portfolios; such limits preventing smaller balls from the extremes of the characteristic distribution dominating. The ball with the highest equally weighted return in period t constitutes the desirable "long" portfolio, whilst that with the lowest is the candidate for the "short" portfolio. Identification of these portfolios is step 2 of the process. Taking a long (short) position on these portfolios, equal- and value-weighted returns are calculated for period $t+1$ as step three. Finally, we test the composition of the firms within the two portfolio balls using a Kolmogrov-Smirnov test.

Results for the TDABM portfolios are dependent on the ball radius. Consideration of mutliple radii allows potential selection of that radius which offers the highest return. Radius selection is taken as that which offers the highest evenly weighted portfolio return at time t. In each case 1000 iterations of the TDABM algorithm

[13] It is noted that the short run reversal may make the use of the highest and lowest current period returns suboptimal. The positive abnormal returns to the strategies suggest that in practice selection of the highest returning ball to long is more optimal than the selection of the lowest.

—- Figure 8 Here —-

are performed (where different ϵ-nets are used to construct the TDABM graphs). Our selected portfolio is then that which offers the highest return at time t defined by radius and iteration number. This information is then entirely observable by the investor looking to form their portfolio for month $t + 1$

Figure 8 informs how the changing radius of the TDABM graph impacts the returns gained on the long-short strategy. As the radius increases so the disparity between the highest and lowest returning balls shrinks, manifesting as the downward slope in panel (a). Interestingly, when considering the subsequent period, the smaller radii offer a good guide; positive returns to the long short period in $t + 1$ are noted. There is then a range over which next period returns are negative on average before they approach zero from a radius of 0.5 upwards. None of these are significant as evidenced by the wide percentile lines on panel (b). However the strategy is not to choose one realisation at random, but to choose the portfolio which offers the highest in sample return. Panels (c) and (d) chart the properties of those portfolios. Solid lines indicate the long portfolio. We see very different behaviour between a radius of 0.2 and around 0.35, but generally the long portfolio contains much larger firms with medium to low book-to-market ratios. The short portfolio is always smaller firms and has lower book-to-market ratios. Around radius $\epsilon = 0.65$ the short portfolio becomes the one with the higher book-to-market ratio. Comparing across the four panels radius matters. Because the strategy that follows allows investors to choose the radius they see giving the highest in sample return it follows that lower radii will be selected.

6.2. *Common Anomalies Examples*

TDABM allows similar segmentation of the seven variable space to that performed for the bivariate example. Again only balls with 30 stocks or higher are considered and the investor is assumed to take a long (short) position on the ball with the highest (lowest) in sample return. For each radius 1000 iterations of the BM algorithm are used. Because there are more axes we no longer consider radii below 0.10 as they

—- Figure 9 Here —-

generate low numbers of portfolios with 30 or more stocks contained within.

Figure 9, based on June 2018 data, presents a similar story to the bivariate case, in sample returns falling as the radius rises and the month ahead returns being higher for lower radii but not significant. Results for 1978, 1988, 1998 and 2018 are similar, but are omitted for brevity[14]. In panels (c) to (i) the values of the seven characteristics in the long and short portfolios are plotted as solid and dotted lines respectively. The long has larger more profitable firms with higher earnings-to-price and dividend yield. BM, investment and cashflow are less consistent. These results are not suggestive that these ratios are less informative on returns, but investors may note these inconsistencies when planning portfolios of varying size. Panel (j) adds the returns of the long and short portfolios that are selected for reference. This is the pair at each radius that offer the largest differential. Note here the variability at low radii in the short portfolio; this then appears as the bumps in the upper line of panel (a). Behind this lies the limit of 30 as a minimum ball size because at these low radii many balls are close to that 30 threshold.

As with the TDABM visualisations, these graphs are only a snapshot for one month. Of more value will be the properties of the portfolios over time. These are explored in the subsequent discussion of mispricing.

6.3. *Bivariate Portfolios and Mispricing*

First consider the use of only firm size and book-to-market ratio for portfolio construction. Following Fama and French (1992), Fama and French (1993) and subsequent works it is understood that the highest return is on the smallest stocks and those with the higher book to market ratios. Table 4 provides summary statistics on the composition of the two portfolios. We continue to only consider balls with at least 30 stocks and use the best long-short return from 1000 repetitions of the BM algorithm.

—- Table 4 Here —-

[14] These results are available on request from the authors as Appendix G.

To understand potential mispricing of the considered strategies nine common factor models are employed. These are the capital asset pricing model (CAPM) after Lintner (1965), Sharpe (1964), and Treynor (1962), the FF3 of Fama and French (1993), the FF3 augmented with momentum after Carhart (1997) (FF4), the Fama and French (2015) five factor model (FF5) and associated augmentation with momentum (FF6), the mispricing factors model of Stambaugh and Yuan (2017) (STY), the q-factor model of Hou et al. (2015) (HKZ), and the recently published time factors of Daniel et al. (2020) (DHS). Existence of significant abnormal returns against these nine may be taken as evidence that the portfolio strategy is mispriced. Factor data is taken from the respective author websites[15].

—- Table 5 Here —-

For comparison with our TDABM portfolios Table 5 presents the returns from longing the smallest stocks and shorting the largest, longing the highest book-to-market ratio and shorting the smallest, and the joint sort of longing the smallest and highest book-to-market, whilst shorting the largest and lowest book-to-market. Further comparison comes from selecting the highest (lowest) current returns of the 5x5 sorted portfolios as long (short) candidates. TDABM gives us the ball with the highest current period return as a long portfolio, and the ball with the lowest current period return as a potential short portfolio.

A consistent message from Table 6 is that it is possible to generate alpha using the TDABM approach across two-dimensions. Panel (a) uses monthly rebalancing, whilst panel (b) assumes that portfolios are only updated at the end of each June per Fama and French (1993). Under equal weighting panel (a) informs that the monthly rebalancing generates an alpha of around 1.5% and that this is far in excess of the approximately 1% generated by the equivalent strategy from the 5x5 bivariate sort.

—- Table 6 Here —

[15] Factor data for the FFC6 is taken from the website of Ken French: `http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/`, the Hou et al. (2015) factors are from global-q: `http://global-q.org/index.html`, Mispricing factors of Stambaugh and Yuan (2017) are taken from Robert Stambaugh: `http://finance.wharton.upenn.edu/stambaug/` and (Daniel et al. 2020) from `https://sites.google.com/view/linsunhome/`.

In turn both approaches generate a higher alpha than either the univariate sorts, or the theoretical best combination from the 5x5 bivariate sort. Under value weighting similar results emerge, although the magnitude of the alphas is smaller, and the book-to-market univariate sort performs better, it is still the TDABM that offers the highest alpha. These results also hold for the Fama and French (1993), Fama and French (2015) and Stambaugh and Yuan (2017) models[16].

Compared to the understood turnover of approximately 30% on the quintile sorted portfolios, the turnover on TDABM is almost 100%. Consequently, the alphas generated may not be achievable by many. Trading costs of 1bp per leg per 1% turnover as suggested by Chen and Velikov (2017) would reduce the alpha advantage of TDABM, attracting 100bps deduction per leg compared to the BM 5-1 strategy with costs of 30bps. For two of the model weighting combinations BM 5-1 would offer a higher return than the TDABM in the presence of these trading costs. In reality trading costs for portfolio managers are likely to be much lower, and we have already ruled out trading in the extremes of the distribution of stocks by imposing the limit on the minimum ball size; the BM 5-1 continues to trade stocks in both tails of the book-to-market distribution. Indeed costs of 10bps per leg are more commonly applied in the contemporary ML asset pricing literature. It is thus suggested that the TDABM approach has potential to produce investable strategies.

Results in panel (a) of Table 6 follow monthly portfolio rebalancing. Trading costs may be greatly reduced using annual portfolio updating rendering it beneficial to contrast our results with the case that the portfolios are held for one year. Following Fama and French (1993) we update the portfolios at the end of June and hold the stocks from the start of July until the end of the following June. Panel (b) shows that there are then few significant alphas to be had from the TDABM strategy. Rather the assumed relationships from the quintile portfolios deliver the best returns. Longing small firms with high BM ratios, and shorting larger firms with low BM, yields significant alphas.

[16] Results including the alphas for all of the factor models are available on request from the authors.

—- Table 7 Here —-

6.4. *Common Anomalies Portfolios*

When considering the seven common anomalies BM remains a model free means to segment the space, whilst sorted portfolios no longer possess the ability to divide stocks into meaningfully large groupings. For a meaningful comparison a return is made to the work of Liao et al. (2008) and the potential of k-means clustering to produce segmentation. Snow (2020) and Ge et al. (2021) take steps to optimise the k-means clustering for portfolio creation, but in this case we again appeal to the basic premise of the methodology for a comparison with the naive TDABM strategy. Numbers of clusters are guided according to the optimal gap statistic (Tibshirani et al. 2001), and the comparison numbers of balls in the optimal TDABM. In each of the example Junes more than 200 clusters are suggested to be optimal, but this is impractical when the aim is for portfolios with more than 30 stocks contained within. Therefore 100 and 200 clusters are used.

Table 7 reports summary statistics for the properties of the 7 common anomalies considered within this paper for the TDABM and k-means portfolios. In almost all cases paired t-tests inform that the characteristics of the long and short portfolios are significantly different, only the earnings-to-price and dividend yield do not show significant differences. Rank columns inform that there is consistency amongst the selections of TDABM and k-means and that both pick portfolios which align with theory more often than not. The time series of factors in Fama and French (2020) and elsewhere are also consistent with the idea that the sorted portfolios do not consistently yield positive excess returns. Table 7 is therefore aligned with theory and reassuring of the power of TDABM to select profitable portfolios.

Excess returns are reported in the first row of Table 8 and show that the k-means values under value weighting are by far the highest of the six considered. Again it is the same set of nine models that are used for evaluation. However, only the TDABM values have a Newey-West adjusted t statistic greater than 2. This pattern then repeats through the alphas with the largest values again coming under k-means with value

—- Table 8 Here —-

weighting, but none of these being statistically significant. Compared to the bivariate case the TDABM annualised alphas have fallen from around 18% to around 15%, but the significance remains against all but the Daniel et al. (2020) factor model. Considering the Sharpe ratios also finds in favour of TDABM since the annualised value here is 1.7 compared to a 1.3 maximum for k-means. As in the bivariate case, the turnover from TDABM is close to 100% each month and likewise for k-means a similarly high turnover is noted. As with the bivaraite case approaches that optimise selection rather than adopting long-standing rules fare badly in annual reconstitution. For brevity therefore the annual comparison is not included.

It is noted here that there are alternative clustering techniques that may be contrasted with TDABM, and that the RF tree approach of Bryzgalova et al. (2020) may form a suitable comparison also. However this section has focused on unsupervised alternatives where only the most recent return was used as a guide to what may happen in the next period. Allowing all of the models to train on past data would be a fruitful exercise and, based on the first steps taken in this paper, represent potential in identifying mispriced portfolios.

6.5. *Summary*

TDABM offers potential to generate alpha even after the deduction of trading costs. However, the level is below that being generated by other segmentations of the characteristic space. Critically, where approaches such as the application of regression trees to generate splits from historic data, the TDABM approach discussed here uses only contemporary information and is model free. Extending the analysis to incorporate optimal choices of radius from historic next month returns, for example, offers another direction in which the application of TDABM within finance could be developed.

7. Conclusions

This simple exposition has sought to introduce the TDA Ball Mapper algorithm of Dłotko (2019) as a means of understanding the relationships, and interdependencies, of firm characteristics in the cross-section of stock returns. Further we have demonstrated how inference upon the generated plots offers potential in understanding the cross section of stock returns. Using a two axis example it was shown that, in line with the inference from recent ML literature, links are not linear in the way that the established literature has often assumed. Extending to a seven dimensional set based upon commonly studied stock return determinants, the value of the TDA approach in unpacking the correlations between financial ratios and stock returns was again demonstrated. Because BM does not inform on causality, only the evidence of possible motivation for observed return patterns is statistically justifiable as a conclusion. Whether looking at bivariate or multi-dimensional data a deeper appreciation of what this extensive period of financial data relays about the market is delivered by our unique application of TDA.

BM offers further strengths in then understanding how well cross-sectional models are pricing stocks. Our second contribution was to show the limited differentials between the residuals emerging from OLS and ML models. For five months it was seen how models trained on the previous 11 months of data produced very similar predictions for the following month's returns. Typically ML models are trained on a far broader set of characteristics, and they are offered a much longer training set than the eleven months given here. Considering the inference that could come from TDABM on the back of such well trained models would be a valuable next step based on the formative illustrations here.

Utilising TDABM diagrams as a means of portfolio construction we demonstrated how a long-short strategy on the balls with maximum and minimum present returns can generate significant alpha in the subsequent month. In the bivariate case the models strongly outperformed bivariate sorted portfolios. In the multivariate case a comparison with clusters formed using the unsupervised k-means algorithm also showed a superiority for TDABM. Again our aim here was to generate a discussion of the potential of TDABM in a future research agenda; properly trained models have natural potential to beat those presented here. Nonetheless the potential to generate alpha cannot be ignored.

Investigation of the selection of anomalies in this paper is not exhaustive, our objective being to illustrate BM as an approach. Development of this analysis would be highly fruitful. A further possibility that is not explored here is to first reduce the set of characteristics using either principal component analysis or ML in the spirit of Gu et al. (2020) and others. Applying TDABM onto the reduced set is then likely to offer further insight into the second stage regressions that the ML literature papers perform on their reduced sets.

This paper thus offers a new understanding of the link between established firm financials and stock returns. Using a robust and model free methodology that is built from the data, and faithfully respects the topology thereof, pre-assumed monotonic relationships are shown absent in multiple dimensions. Early promise for investors has been identified and it is for subsequent work to build on that strong foundation. Future research on the fuller set of characteristics offers potential, as does the developing ability of TDA Ball Mapper to work in data reduction, in dynamic settings and in conjunction with ML. A rich seam of exploration awaits. For investors, analysts and those seeking a deeper understanding of markets it is an invaluable tool to undergird effective organisation is provided. The full power of the tool is there to be exploited.

Acknowledgements

P.D. acknowledges the support of the Dioscuri program initiated by the Max Planck Society, jointly managed with the National Science Centre (Poland), and mutually funded by the Polish Ministry of Science and Higher Education and the German Federal Ministry of Education and Research.

References

Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician 27*(1), 17–21.

Arnott, R., C. R. Harvey, and H. Markowitz (2019). A backtesting protocol in the era of machine learning. *The Journal of Financial Data Science 1*(1), 64–74.

Basu, S. (1977). Investment performance of common stocks in relation to their price-earnings ratios: A test of the efficient market hypothesis. *The Journal of Finance 32*(3), 663–682.

Basu, S. (1983). The relationship between earnings' yield, market value and return for nyse common stocks: Further evidence. *Journal of Financial Economics 12*(1), 129–156.

Black, F. and M. Scholes (1974). The effects of dividend yield and dividend policy on common stock prices and returns. *Journal of Financial Economics 1*(1), 1–22.

Breiman, L., J. Friedman, C. J. Stone, and R. A. Olshen (1984). *Classification and regression trees.* CRC press.

Brooks, C., A. G. Hoepner, D. McMillan, A. Vivian, and C. Wese Simen (2019). Financial data science: the birth of a new financial research paradigm complementing econometrics? *European Journal of Finance 25*(17), 1627–1636.

Bryzgalova, S., M. Pelger, and J. Zhu (2020). Forest through the trees: Building cross-sections of stock returns. *Available at SSRN 3493458*.

Carhart, M. M. (1997). On persistence in mutual fund performance. *The Journal of finance 52*(1), 57–82.

Carlsson, G. (2009). Topology and data. *Bulletin of the American Mathematical Society 46*(2), 255–308.

Carriere, M. and S. Oudot (2018). Structure and stability of the one-dimensional mapper. *Foundations of Computational Mathematics 18*(6), 1333–1396.

Chen, A. Y. and M. Velikov (2017). Accounting for the anomaly zoo: A trading cost perspective. *Available at SSRN, 3073681*.

Daniel, K., D. Hirshleifer, and L. Sun (2020). Short-and long-horizon behavioral factors. *The Review of Financial Studies 33*(4), 1673–1736.

Dłotko, P. (2019). Ball mapper: a shape summary for topological data analysis. *arXiv preprint arXiv:1901.07410*.

Dlotko, P. (2019). *BallMapper: Create a Ball Mapper graph of the input data.* R package version 0.1.0.

Fama, E. F. and K. R. French (1988). Dividend yields and expected stock returns. *Journal of Financial Economics 22*(1), 3–25.

Fama, E. F. and K. R. French (1992). The cross-section of expected stock returns. *the Journal of Finance 47*(2), 427–465.

Fama, E. F. and K. R. French (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics 33*(1), 3–56.

Fama, E. F. and K. R. French (2015). A five-factor asset pricing model. *Journal of Financial Economics 116*(1), 1–22.

Fama, E. F. and K. R. French (2020). Comparing cross-section and time-series factor models. *The Review of Financial Studies 33*(5), 1891–1926.

Fama, E. F. and J. MacBeth (1973). Risk, return, and equilibrium: Empirical tests. *Journal of Political Economy 81*, 607–636.

Feng, G., S. Giglio, and D. Xiu (2020). Taming the factor zoo: A test of new factors. *The Journal of Finance 75*(3), 1327–1370.

Fisher, J. D., D. W. Puelz, C. M. Carvalho, et al. (2020). Monotonic effects of characteristics on returns. *Annals of Applied Statistics 14*(4), 1622–1650.

Ge, S., S. Li, and O. Linton (2021). Dynamic peer groups of arbitrage characteristics.

Green, J., J. R. Hand, and X. F. Zhang (2017). The characteristics that provide independent information about average us monthly stock returns. *The Review of Financial Studies 30*(12), 4389–4436.

Gu, S., B. Kelly, and D. Xiu (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies 33*(5), 2223–2273.

Hartigan, J. A. and M. A. Wong (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics) 28*(1), 100–108.

Harvey, C. R., Y. Liu, and H. Zhu (2016). . . . and the cross-section of expected returns. *The Review of Financial Studies 29*(1), 5–68.

Haugen, R. A., N. L. Baker, et al. (1996). Commonality in the determinants of expected stock returns. *Journal of Financial Economics 41*(3), 401–439.

Hoepner, A. G., D. McMillan, A. Vivian, and C. Wese Simen (2021). Significance, relevance and explainability in the machine learning age: an econometrics and financial data science perspective. *The European Journal of Finance 27*(1-2), 1–7.

Hou, K., C. Xue, and L. Zhang (2015). Digesting anomalies: An investment approach. *The Review of Financial Studies 28*(3), 650–705.

Kozak, S., S. Nagel, and S. Santosh (2019). Shrinking the cross-section. *Journal of Financial Economics*.

Lettau, M. and M. Pelger (2020). Factors that fit the time series and cross-section of stock returns. *The Review of Financial Studies 33*(5), 2274–2325.

Liao, S.-H., H.-h. Ho, and H.-w. Lin (2008). Mining stock category association and cluster on taiwan stock market. *Expert Systems with Applications 35*(1-2), 19–29.

Lintner, J. (1965). Security prices, risk, and maximal gains from diversification. *The Journal of Finance 20*(4), 587–615.

Masters, T. (1993). *Practical neural network recipes in C++*. Morgan Kaufmann.

Matejka, J. and G. Fitzmaurice (2017). Same stats, different graphs: generating datasets with varied appearance and identical statistics through simulated annealing. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, pp. 1290–1294.

Newey, W. K. and K. D. West (1987). Hypothesis testing with efficient method of moments estimation. *International Economic Review*, 777–787.

Ou, J. A. and S. H. Penman (1989). Financial statement analysis and the prediction of stock returns. *Journal of Accounting and Economics 11*(4), 295–329.

Patton, A. J. and A. Timmermann (2010). Monotonicity in asset returns: New tests with applications to the term structure, the capm, and portfolio sorts. *Journal of Financial Economics 98*(3), 605–625.

Pukthuanthong, K., R. Roll, and A. Subrahmanyam (2019). A protocol for factor identification. *The Review of Financial Studies 32*(4), 1573–1607.

Romano, J. P. and M. Wolf (2013). Testing for monotonicity in expected asset returns. *Journal of Empirical Finance 23*, 93–116.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The Journal of Finance 19*(3), 425–442.

Shively, T. S., T. W. Sager, and S. G. Walker (2009). A bayesian approach to non-parametric monotone function estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71*(1), 159–175.

Snow, D. (2020). Machine learning in asset management—part 1: Portfolio construction—trading strategies. *The Journal of Financial Data Science 2*(1), 10–23.

Stambaugh, R. F. and Y. Yuan (2017). Mispricing factors. *The Review of Financial Studies 30*(4), 1270–1315.

Tibshirani, R., G. Walther, and T. Hastie (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63*(2), 411–423.

Treynor, J. L. (1962). Jack Treynor's' toward a theory of market value of risky assets'. *Available at SSRN 628187*.