

GROWTH & VALUE STOCK INDICES UNDER THE EDA MICROSCOPE

ABSTRACT

In this paper, the author uses geometrical and topological aspects of Exploratory Data Analysis (EDA) to examine Standard and Poors (S&P), MSCI and Thomson Reuters Indices' ways of determining which stocks are growth and which are value.

The results of the analysis are that two of the firms – S&P and Thomson Reuters Indices - do a very good job of determining what S&P calls “pure growth” and “pure value” stocks. This is found to be true to a large extent because their respective point cloud data tends to fall into linear clusters and hence is amenable to their respective linear separation techniques. A lack of linearity in MSCI's point cloud data is more often than not foiling their linear methodologies so “the jury is still out” on MSCI's ability to affectively separate growth and values stocks. The techniques used to arrive at these conclusions include the use of ggobi (a “grand tour” data visualization system), isomaps (a nonlinear data reduction tool), model-based clustering and multiresolution bootstrap resampling.

The paper also shows that the “core” designation that has been put forward by such firms as Lipper and Morningstar and individuals such as Ron Surz is a true classification, at least if one uses Thomson Reuters Indices' fundamental factors to separate growth and value. As to correctly identifying the value effect, i.e., those times when the Fama-French value benchmark outperforms its growth benchmark, only Thomson Reuters' growth and value indices consistently demonstrate the value effect

Introduction

U.S. Growth and value stock indices have had been around for approximately 17 years. According to a Barra document: “In 1992, Standard and Poors and Barra began a collaboration to produce Growth

and Value subsets of S&P's industry-leading equity indexes. Academic research

pioneered by Nobel Laureate William Sharpe, and continued by Eugene Fama, Kenneth French and others have confirmed the validity of the growth/value distinction in terms of differential returns over time. The sole criterion for the S&P/Barra Growth/Value split is the book value of a common equity divided by the market capitalization of a firm.”¹ The S&P/Barra indices were the first publically available growth and value stock indices and were soon followed by MSCI and then Morningstar and Lipper who build growth, value and core mutual fund indices.

Since that time, the number of providers of growth and value indices has grown and methodologies have changed. The change in methodologies has come about primarily because of the influence of papers by Fama and French [1992, 1996]. Several other economists have done research in this area subsequently, for example Chan, Karceski and Lakonishok [2000] find that value stocks do have on average better returns in the U.S. and Chan and Lakonishok [2004] show that this effect applies to countries outside the U.S. as well.

The empirical research has shown that in a cross-section of stocks returns, investor returns often increases with book-to-market ratio. The reason for this is in dispute. One side argues that the market is inefficient or possibly irrational; another that value stocks (those with higher book-market ratios) are riskier investments. Statistical analysis of historical data has not resolved the issue and various theoretical models have been proposed with mixed success (please see Stutzer [2003] for a good survey as well as a response on this topic).

The growing recognition of Fama and French's work has meant that multifactor fundamental models have become very popular. MSCI Barra has built a

¹ Barra, <http://www.barra.com/research/Description.aspx>

significant business out of building multifactor models and many other firms have followed suit. And mention needs to be made of the grandfather of all this research, the Arbitrage Pricing Theory (APT) of Roll and Ross [1980].

With adoption of multifactor models has come the increasing use of fundamental factors to distinguish growth and value stocks by index providers. It has only been in the current decade that two of the biggest providers of growth and value indices – S&P and MSCI – have adopted multifactor techniques to distinguish growth and value stocks. What this paper will show is that the factors each provider uses to distinguish growth and value, is successful (or not successful) based upon the correct understanding of the topology of the data's multidimensional point clouds. This knowledge is key to evaluating the successes and drawbacks of each provider's methodology. The paper will make clear that none of the providers are fully exploiting the geometry and related topology of the point cloud data they are collecting and in some cases are making methodological assumptions that not justified by the topology of the data.

The paper will show that S&P's point cloud topology more often than not generates two almost linear clusters at an angle to each other so S&P is able to correctly classify growth and value stocks using its combination of linear measures. MSCI's point cloud topology is much more difficult to work with and with no clear geometry (distance metric) to be seen. This, the author thinks, are the reasons for the "not as good" performance of the MSCI classification scheme. Thomson Reuters takes an approach similar to Fama and French, i.e., they use cross-sectional robust regression to separate growth and value, and their data's point clouds breaks into two linear clusters similar to S&P's and hence their linear regression works in their favor as well.

Section 2 of the paper will review the construction methodologies of S&P, MSCI and Thomson Reuters Indices. Section 3 will analyze the geometry of the data which the study uses (U.S. stock fundamentals from 1984 through 2008). Section 4 will analyze the data using EDA. And Section 5 will have the conclusions.

Section 2: Index Provider Methodologies

As is common in the passive or benchmark index business, construction methodologies tend to be very transparent. Each firm's website will normally have a pdf freely available that details how each benchmark index is constructed. These websites are provided in the first reference of each firm's methodology.

Section 2.1: S&P

S&P U.S. growth and value methodology is straight forward and easy to replicate.² In the document entitled *S&P U.S. Style Indices, Index Methodology*, S&P writes:

Standard & Poor's U.S. Style index solutions address two distinct needs. The first is for exhaustive style indices that can effectively form the basis for index funds and derivatives, providing broad, cost-efficient exposure to a certain style segment. The second need is for narrow, style-pure indices that provide a pure style return series, and serve as the basis for style-concentrated investment vehicles or "style spread" products.

With the S&P U.S. Style indices, Standard & Poor's is providing a comprehensive Style index solution by building separate style and pure-style indices, and by making available a consistent set of stock-level Style Scores and Style indices.

The **Style** index series divides the complete market capitalization of each parent index approximately equally into growth and value indices. This series covers all stocks in the parent index universe, and uses the conventional, cost-efficient market cap-weighting scheme. Stocks that do not fall into Pure Style baskets have their market caps distributed between growth and value indices.

The **Pure Style** index series identifies approximately one third of the parent index's market capitalization as Pure Growth and one third as Pure Value. There are no overlapping stocks, and these indices do not have the size

² The website for their U.S. index methodology is: http://www2.standardandpoors.com/portal/site/sp/en/us/page.family/indices_ei_us/2,3,2,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0.html

bias induced by market capitalization weighting. Rather, stocks are weighted in proportion to their relative style attractiveness.

S&P's list of factors to determine value are: dividend yield, sales-to-price ratio, cashflow-to-price ratio and book-price ratio. For growth, S&P uses 5 year EPS growth rate, 5 year sales-per-share growth rate, and 5 year internal growth rate (defined as return on equity x earnings retention rate). No mention is made of the required coverage for each factor (something MSCI does mention). In reference to this point, this paper will show that lack of coverage for any one value stock factor is not significant enough to misclassify pure value stocks but missing a single growth factor can cause a significant jump in the number of misclassifications of pure growth stocks.

S&P goes on in their document to say how it constructs style scores:

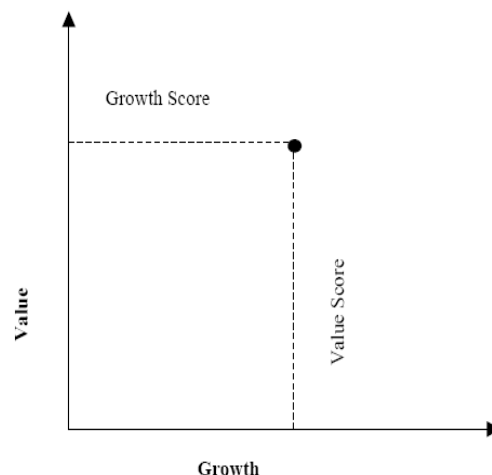
Style Scores. Raw values for each of the above factors are calculated for each company in the S&P/Citigroup Broad Market Index (BMI) universe, which has approximately twice as many stocks as the S&P Composite 1500. These raw values are then standardized by dividing the difference between each stock's raw score and the mean of the entire set by the standard deviation of the entire set. A Growth Score for each company is computed as the average of the standardized values of the three growth factors. Similarly, a Value Score for each company is computed as the average of the standardized values of the four value factors.

The simple averaging process assumes each factor is equally important. Different factors will clearly have different discriminating powers over time, but the equal weighting approach is chosen to meet the design goal of simplicity.

At the end of this step each stock has a Growth Score and a Value Score, as shown below, with

growth and value being measured along separate dimensions.

Exhibit 1: Measuring Growth and Value Along Separate Dimensions



For Stock X,

$G_{i,x}$ = Standardized value of Growth Factor I for stock X, I=1 to 3.

$V_{j,x}$ = Standardized value of Value Factor J for stock X, J=1 to 4.

SG_x = Growth Score of X = $1/3 (G_{1,x} + G_{2,x} + G_{3,x})$

SV_x = Value Score of X = $1/4 (V_{1,x} + V_{2,x} + V_{3,x} + V_{4,x})$

S&P's use of standardization could point to a potential weakness in their methodology, especially if the dataset is plagued by outliers. As it turns out (see Section 3), outliers are and are not a significant issue for S&P's data.

Finally, the S&P manual says that to construct indices after the average z-scores are computed:

Stocks within each parent index are ranked based on growth and value scores. A stock with a high Growth Score would have a higher Growth Rank, while a stock with a low Value Score would have a lower Value Rank. (For example, the S&P MidCap 400 constituent with the highest Value Score would have a Value Rank of 1, while the constituent with the lowest would have a Value Rank of 400.)

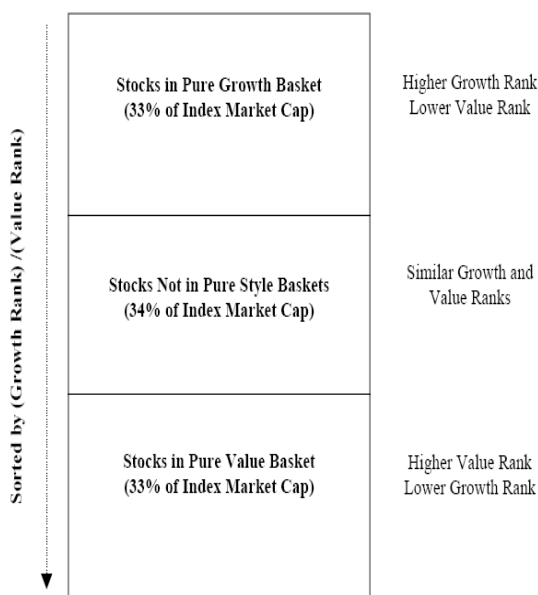
The index constituents are then sorted in ascending order of their Growth Rank/Value Rank. The stocks

at the top of the list have a higher Growth Rank (or high Growth Score) and a lower Value Rank (or low Value Score) and, therefore, exhibit pure growth characteristics. The stocks at the top of the list, comprising 33% of the total market capitalization of the index, are designated as the Pure Growth basket.

The stocks at the bottom of the list have a higher Value Rank (and Value Score) and a lower Growth Rank (and Growth Score) and, therefore, exhibit pure value characteristics. The stocks at the bottom of the list, comprising 33% of the total market capitalization of the index, are designated the Pure Value basket.

The stocks in the middle of the list have neither pure growth nor pure value characteristics. The distribution of the index universe into pure style regions is illustrated below.

Exhibit 2: Pure Style Baskets



Finally, to classify those stocks that are not pure growth or pure value, S&P uses the Euclidean distance in both 1 and 2D to distinguish non-pure growth and non-pure value stocks³.

³ See pg. 10 and Appendix I, *op. cit.* for more details.

Section 2.2: MSCI

Similar to S&P, MSCI takes a multi-factor approach to determining growth and value stocks. Their factors are:

For Value:

- Book value to price ratio (BV / P)
- 12-month forward earnings to price ratio (E fwd / P)
- Dividend yield (D / P)

And For Growth:

- Long-term forward earnings per share (EPS) growth rate (LT fwd EPS G)
- Short-term forward EPS growth rate (ST fwd EPS G)
- Current Internal Growth Rate (g)
- Long-term historical EPS growth trend (LT his EPS G)
- Long-term historical sales per share (SPS) growth trend (LT his SPS G)

And like S&P, MSCI computes value and growth z-scores for stock. However, unlike S&P, prior to computing z-scores, MSCI winsorizes each of the eight variables at $\bullet = 5\%$. This is to control for outliers, a topic we will return to in Section 3.

MSCI's computation of value z-scores is the same as S&P's, i.e., the simple average of the available z-scores. So if one value variable is missing, the value z-score is the average of the two. For growth z-scores, MSCI, in the author's view, rightly takes into account the affect of what a missing variable can mean for the growth z-score. To quote from their document⁴:

Computing the growth z-score differs from computing the value z-score because missing variable z-scores are not excluded from the calculation and their z-scores are set to zero (i.e., to the average of the market capitalization index). This is because variables used to define growth investment style characteristics are less correlated to one another compared to those that are used to define value investment style characteristics. Hence, excluding missing variables from the growth z-score calculation could result in a biased growth z-score that is influenced too significantly by variable z-

⁴ MSCI Methodology Book: US Equity Indices, November 11, 2008, pg 21 and 22.

scores that are not missing. In addition, this treatment ensures that in cases where many variables are missing, the resulting growth z-score is close to the market average.

The growth z-score is calculated as follows:

$$\text{Growth Z-Score} = \frac{1}{6} (2 * Z_{LT_fnd_EPS_G} + Z_{ST_fnd_EPS_G} + Z_g + Z_{LT_htc_EPS_G} + Z_{LT_htc_SPS_G})$$

For instance, if the long-term forward EPS growth rate variable is missing:

$$\text{Growth Z-Score} = \frac{1}{6} (Z_{ST_fnd_EPS_G} + Z_g + Z_{LT_htc_EPS_G} + Z_{LT_htc_SPS_G})$$

For a financial company:

$$\text{Growth Z-Score} = \frac{1}{5} (2 * Z_{LT_fnd_EPS_G} + Z_{ST_fnd_EPS_G} + Z_g + Z_{LT_htc_EPS_G})$$

All stocks also receive a VIF and GIF measure. MSCI uses a squared distance constrained to $[0,1]$ to distinguish VIF and GIF.

$$\text{value contribution} = \frac{\text{value z-score}^2}{\text{distance}^2} = \frac{\text{value z-score}^2}{\text{value z-score}^2 + \text{growth z-score}^2}$$

$$\text{growth contribution} = \frac{\text{growth z-score}^2}{\text{distance}^2} = \frac{\text{growth z-score}^2}{\text{value z-score}^2 + \text{growth z-score}^2}$$

$$\text{value contribution} + \text{growth contribution} = 1$$

Once this computation is complete, the construction of the indices can begin. The final assembly involves with what MSCI calls “distance from the origin.” Distance from the origin is the sum of the squared value z-score and squared growth z-score. The first stocks to go into an index are those stocks with the greatest distance from the origin and either $VIF > GIF$ for value stocks or $GIF > VIF$ for growth stocks. This process continues with the effect being that the “more distant” a stock is from the origin the more likely it is to go into its expected index. But as the distance from the origin becomes smaller, the chances grow that a stock with a bias toward value (growth) will go into the growth (value index). The reasons for this occurring are explained on pages 25 – 28 in the methodology book referenced in footnote 4 above.

Section 2.3: Thomson Reuters Indices

The Thomson Reuters Indices method to compute growth and value (as well as core in this case) is both similar and yet very different from MSCI and S&P. Thomson Reuters Indices uses a multi-factor model but uses robust regression techniques to

determine coefficients that will be used to rank stocks.

Thomson Reuters Indices’ four factor model, based on Lakonishok and Chan (2004), uses book equity/market equity, cashflow/share, sales/share and earnings/share from $t-1$. These are the independent variables in the regression. The 12 month return for all stocks at t is the dependent variable.

Thomson Reuters Indices then fits a cross-sectional regression to the variables in order to derive coefficients that will be used for year t fundamentals. The current year’s fundamental factors are multiplied by last’s year’s coefficients and the sums of these multiplications are then ranked in descending order. Thomson Reuters then takes the first 30% of the ranked stocks and labels them value, the next 40% and labels those as core, and the final 30% are growth. This ranking is very similar to the way Fama and French determine growth and value in their multi-factor cross-sectional regressions (though they do not label the middle 40% in any way).

Though the regression coefficients for each year are not strictly autonomous, there is typically very little variation in the coefficients from one year to the next. This means that in most years the application of the prior year’s coefficients will produce a list very similar to what would have been produced if the Lakonishok and Chan methodology had been used in year t .

Thomson Reuters Indices’ cross-sectional regression is a robust regression that uses Hampel’s redescending-M estimator. A redescending M estimator says that the generalized maximum likelihood estimation is:

$$\sum_{i=1}^n p(x_i, \theta)$$

where p is a function such that p is everywhere differentiable, at least for its first derivative.

The solutions to the generalized maximum likelihood estimator:

$$\hat{\theta} = \arg \min_{\theta} \left(\sum_{i=1}^n p(x_i, \theta) \right)$$

are called M-estimators. The function p , or its derivative, ψ , can be chosen in such a way that the estimator has desirable properties (in terms of bias and efficiency) when the data are truly from the assumed distribution, and 'not bad' behavior when the data are generated from a model that is, in some sense, close to the assumed distribution.

Now if p has a first derivative, the estimation of $\hat{\theta}$ is much easier. An M-estimator of type ψ -type T is defined via the measure function:

$$\psi : X \times \theta \rightarrow \mathbb{R}^r$$

It maps a probability distribution function F onto X such that $T(F) \in \theta$. Given this, the following vector equation can be solved:

$$\int_X \psi(x, T(F)) dF(x) = 0$$

If the function ψ decreases to zero as $x \rightarrow \pm\infty$, the estimator is called redescending. Such estimators have some additional desirable properties, such as the complete rejection of gross outliers.

Hampel's redescending M estimators have ψ functions which are odd functions and are defined for any x by:

$$\psi(x) = x \text{ when } 0 \leq |x| \leq a \quad \text{central segment}$$

$$\psi(x) = a \operatorname{sgn} x \text{ when } a \leq |x| \leq b \quad \text{high \& low}$$

$$\psi(x) = \frac{a(r - |x|)}{r - b} \text{ when } b \leq |x| \leq r \quad \text{ends}$$

$$\psi(x) = 0 \text{ when } r \leq |x| \quad \text{tails}$$

For many choices of ψ , no closed form solution exists and an iterative approach to computation is required. It is possible to use standard function optimization algorithms, such as Newton-Raphson. However, in most cases iteratively weighted least squares can be performed; Thomson Reuters Indices uses least trimmed squares.

For redescending functions, the solution may not be unique. This issue is particularly relevant in regression problems. Care is needed to ensure that good starting points are chosen. Robust starting points, such as the median as an estimate of location and the median absolute deviation (MAD) are commonly used. Thomson Reuters Indices uses both the median and MAD in their methodology.

Section 3: The Data and Its Geometry

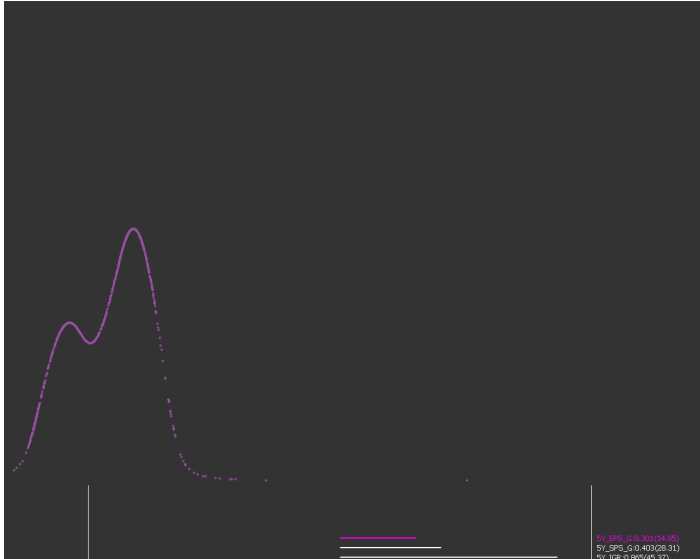
All data used in this paper came from DataScope, in particular, the tool called Reuters Fundamentals. Reuters Fundamentals is a survivorship-bias-free database of U.S. stocks going back to 1984, to 1994 for the remainder of G-7 and the late 1990's for most other countries. There are more than 40,000 firms listed in the database and the database covers more than 100 countries. As the focus of this study is U.S. stocks, we were able to use the longest available historical record in Reuters Fundamentals.

Even though Reuters Fundamentals has 25 years of U.S. history, the data needed to compute one of the MSCI variables - long-term forward earnings per share (EPS) growth rate - was not available across a sufficient number of stocks until 1999. So, data for MSCI data begins in 1999 and ends, like for the others, in 2007/2008. S&P because it requires 5 year long-term EPS and 5 year sales-per-share growth rate, has its first data point in 1988. Thomson Reuters Indices, as it uses the current year's fundamental values, has a starting data of 1984.

All the fundamental data used in this study covers the latest 12 months ending in December of each year. Total returns for the same stocks are computed over the following 12 months and assumes any dividends paid are being re-invested.

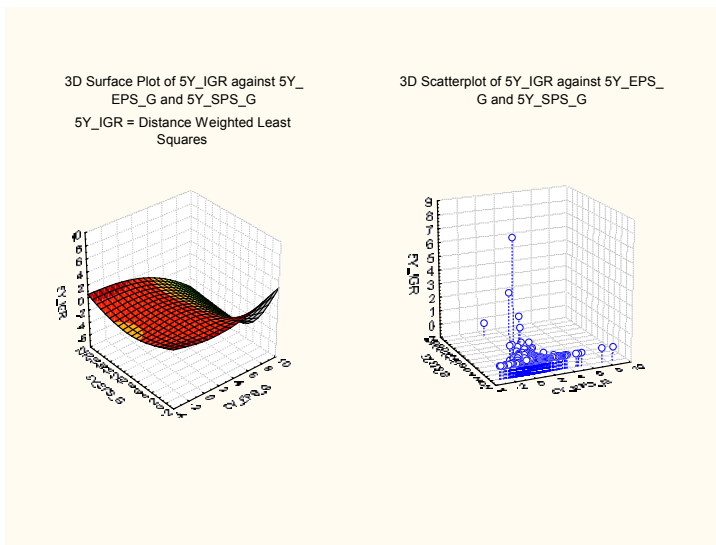
A preliminary view of the S&P data using both ggobi and Statistica® shows a very interesting feature - the multidimensional space that growth and value stocks inhabit is best described as a mixture of distributions. The ggobi plot below shows the bimodal nature of the S&P's fundamentals growth space as of December 1992.

Figure 1



This bimodality does not exist at all times as the 3-D view of the same data in Figure 2 shows:

Figure 2



The manifold for the three growth factors is clearly non-linear but an examination of the same data in scatterplot form shows a significant amount of linearity with the curvature seen in the surface plot seeming to come from a few extreme values and outliers. This appearance of non-linear manifolds being generated by a few extreme values and outliers is repeated across its other factors. The graphical analysis of the linearity of most of the S&P data points to the potential for a good classification of growth and value stocks using linear techniques.

In order to better define the factors complete topology – all seven S&P factors at once – a nonlinear multidimensional scaling (MDS) technique called an isomap is used. For those unfamiliar with isomers, we will develop their theory very briefly. We will follow the development as in Machete (2004) and refer the reader to this text for more detail.

Isomers are a dimensionality reduction technique. It uses a k-nearest neighbor graph to define distances between points, most importantly in this case the inter-point distance matrix.

First two definitions of k-nearest neighbor graphs:

Definition 1:

Given a set $V \subset \mathbb{R}^d$ and a point $v \in V$, define $knn(v)$ to be the set of k-nearest neighbors of v in V . That is $knn(v)$ consists of k points in V closest to v .

Definition 2:

The k-nearest neighbor graph (KNN) on V is defined to be the graph with edge set defined by

$$pq \in E(KNN) \Leftrightarrow p \in knn(q)$$

or $q \in knn(p)$. The mutual k-nearest neighbor graph (MKNN)

requires the points to be the k-nearest neighbors of each other;

$$pq \in E(MKNN) \Leftrightarrow p \in knn(q)$$

$$\text{or } q \in knn(p).$$

Definition 1 states, “That is $knn(v)$ consists of k points in V closest to v .” These are the k points closest to v given a *multidimensional* space V . For S&P data, the space V is seven dimensional as S&P uses seven fundamental factors in its methodology. For those readers more familiar with differential geometry and in particular Riemannian and metric geometries, the “k points in V closest to v ” means that $knn(v)$ will be the geodesic related to those points.

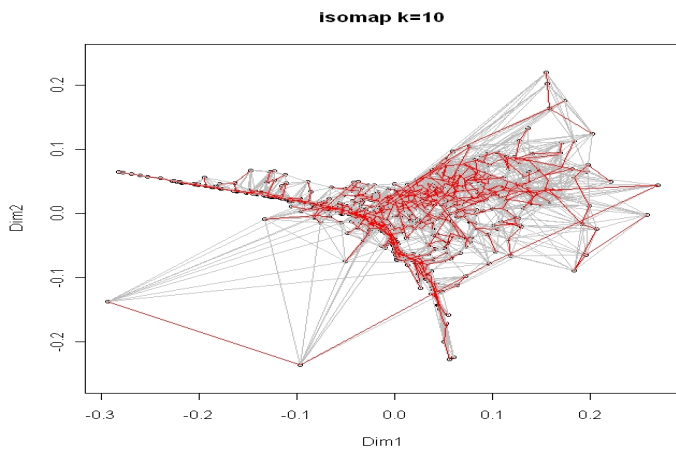
When working with isomers, it is taken as a given that there are a set of points $X = \{x_1, \dots, x_n\}$ with which one first constructs a k-nearest neighbor graph on X . This is used to define the inter-point distance matrix in the following manner:

If x_i and a_x are not in the same graph component then set $d(x_i, x_j) = \infty$. Otherwise define $d(x_i, x_j)$ to be the sum of the lengths of the edges on the shortest path from x_i to a_x . Once the inter-point distance matrix is computed in this way, MDS is applied to reduce the dimensionality. Other techniques could work as well at this point such as clustering or classification algorithms.

In order to determine a minimum k , the analyst looks for the smallest k such that the k -nearest neighbor graph is connected. In our S&P case, $k=10$. Higher values of k can be used to elucidate structure but at some point no further insight is gained from higher values of k .

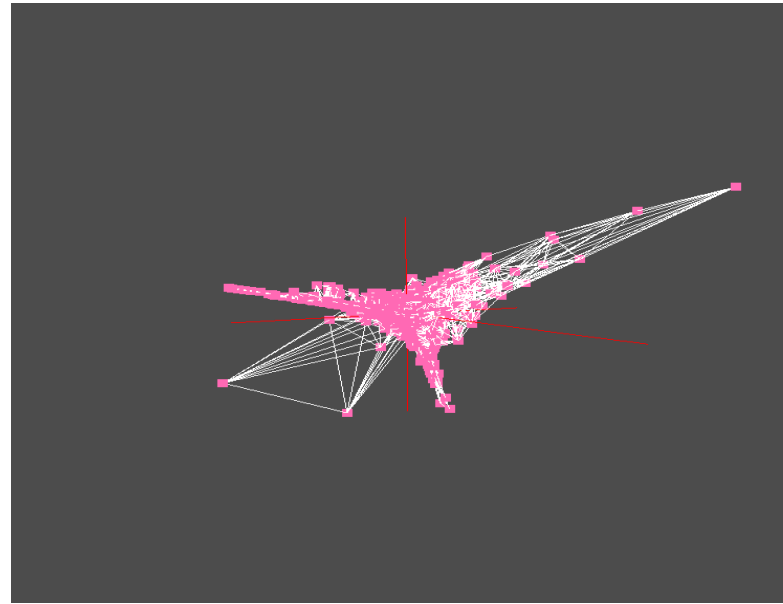
In Figure 3 we show the result of using isomers to reduce the dimensionality of the data from seven to two for S&P data.

Figure 3



The isomer with 10 nearest neighbors shows a curvilinear surface that thins out on both ends and then a loosely connected set of stocks that form a vague, tentacle-like cluster connected to the curvilinear section. Figure 4 (which is a 3D view of the isomer with $k=10$) shows the curvilinear form in 2D is actually two relatively straight lines of some thickness intersecting at an obtuse angle to each other and not oriented along any particular axis.

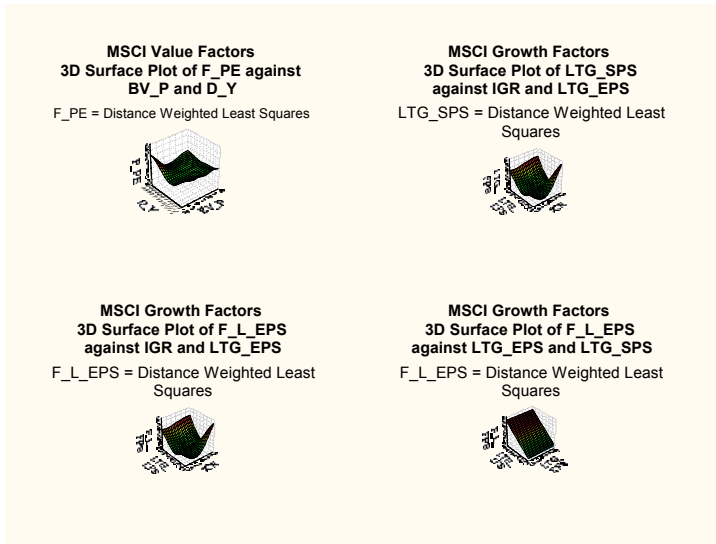
Figure 4



This paper will show that the S&P curvilinear form is composed of growth and value stocks only and that the loosely connected set of stocks behind the curvilinear form will fold into one of these categories. And though there are outliers in the data – more visible in Figure 3 than Figure 4 – their presence will not materially affect the classification outcome.

For MSCI, there is no clear bi- or multimodality visible seen using goby. The 3D plots of the MSCI value factors do however show a very bumpy surface.

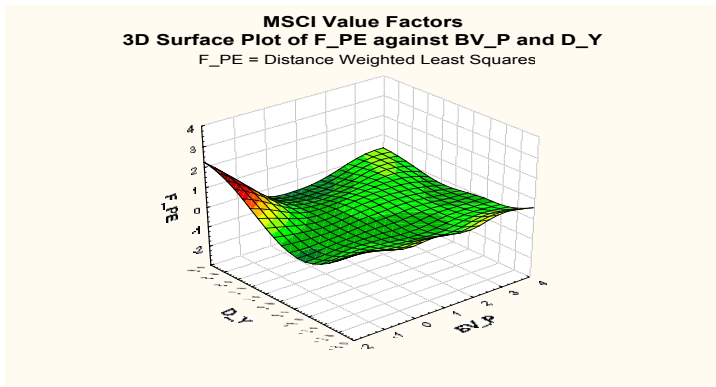
Figure 5



Three of the four surface plot show clear nonlinearities. What little linearity there is, except for the bottom right graph, would at best be local linearity's.

A larger picture of the value graph is below:

Figure 6



A possible function that could fit the MSCI data in Figure 6 is the Fletcher Powell function. The highly multimodal Fletcher and Powell function is a typical representative of nonlinear parameter estimation (regression) problems (such as the distance weighted least squares used to plot the graphs in Figure 6). This function is not symmetric and has its extreme randomly distributed over the surface. The random location of the extreme is achieved by using random matrices $A = (a_{im})$ and $B = (b_{ij})$ in the following way:

$$f(\vec{x}) = \sum_{i=1}^n (A_i - B_i)^2$$

$$A_i = \sum_{j=1}^n (a_{ij} \sin \alpha_j + b_{ij} \cos \alpha_j)$$

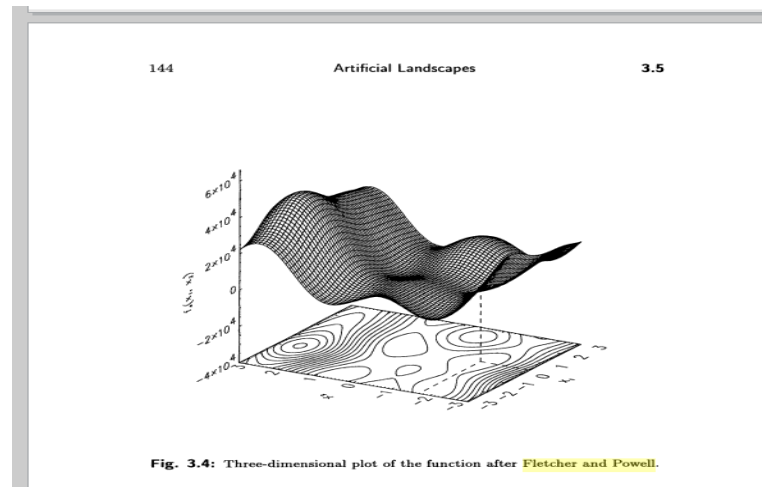
$$B_i = \sum_{j=1}^n (a_{ij} \sin x_j + b_{ij} \cos x_j)$$

$$\vec{x}^* = \vec{\alpha}; f^* = 0; n = 30; -\pi \leq x_i \leq \pi$$

$$a_{ij}, b_{ij} \in [-100, 100]; a_j \in [-\pi, \pi]$$

As Fletcher Powell point out in their article, there are up to 2^n extreme in the interval $|x_i| \leq \pi$. Figure 7 is from Back (1996), page 144 where a good view of a version of the Fletcher and Powell function is available. This graph was computed with matrices A and B being random as well as the vector $\vec{\alpha}$. Altogether there were 1830 points generated ($n = 30$).

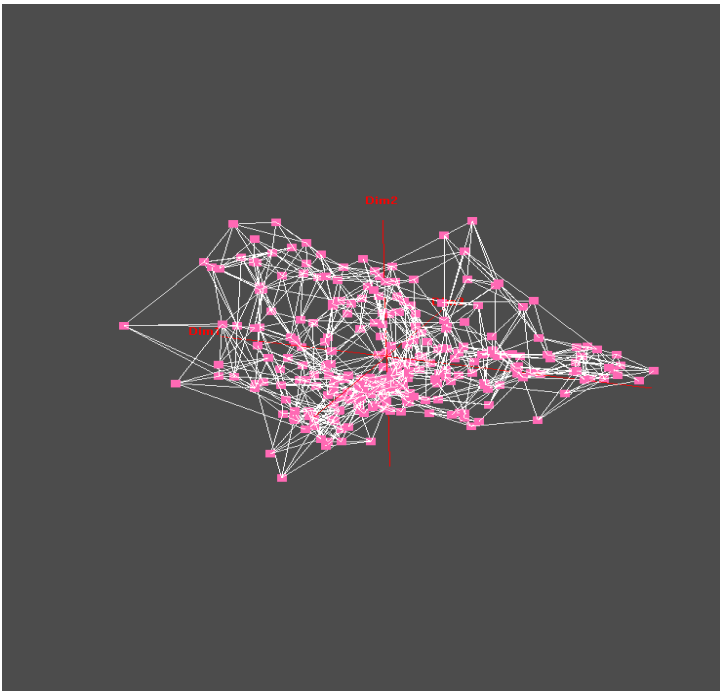
Figure 7



We will come back to Fletcher and Powell as a way of working with Music's nonlinearity in our conclusions.

The 3D isomer of all the MSCI factors does not show any of the linear clustering seen in the S&P 3D isomer. If anything it seems to show that the MSCI factors create a more dispersed view of the stocks with the possibility of some limited clustering:

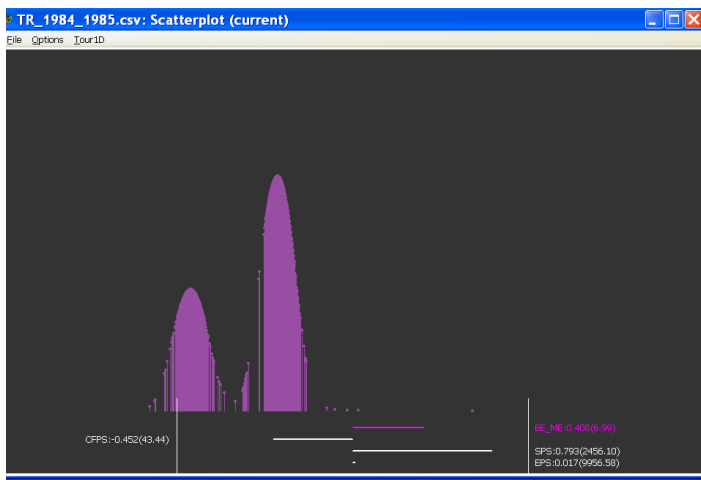
Figure 8



While dinornizes may have helped “clean” this picture, it is difficult to say that there is any evident linear structure in this picture. This point cloud and the related ones for other years, as well as the nonlinearity of the surface plots does not bode well for Music’s linear separation techniques.

The Thomson Reuters Indices data resembles S&P in that it too has bimodal distributions.

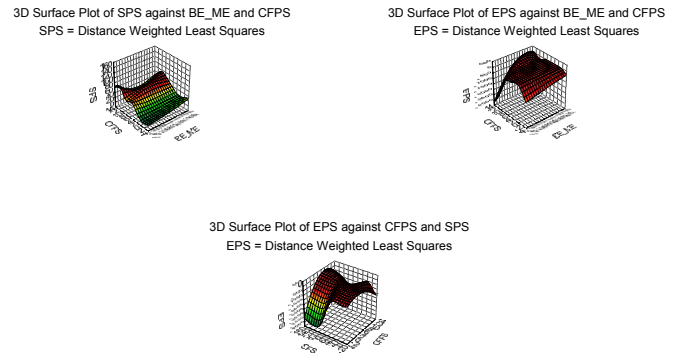
Figure 9



The above is a 1D rotation shot via goby using all four Thomson Reuters Indices factors as of December 1984

Figure 10 shows three of the four factors Thomson Reuters Indices use in its classification scheme: book equity/market equity, sales/share and cash flow/share.

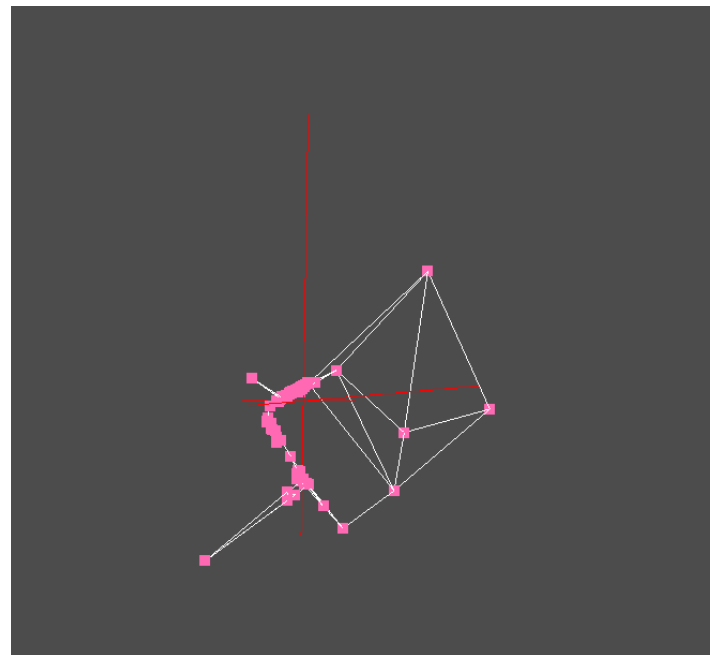
Figure 10



Similar to S&P, much of the curvature in the Thomson Reuters Indices manifolds comes from extreme values and outliers.

Figure 11 is a 3D isomer of the Thomson Reuters Indices data with $k=3$:

Figure 11



The 3D isomer is similar to Sap's in that it too has two distinct linear clusters. This result in conjunction with linearity of much of the data as seen and explained in the surface plots could mean Thomson Reuters Indices, like S&P, will have success in classifying growth and value stocks.

Section 4: Exploratory Data Analysis

In Section 3, we saw linear clusters in the data for S&P and Thomson Reuters Indices. This linearity may help support their respective linear methods of classifying stocks. MSCI data however has no clear linearity which does not bode well for their linear methodology. We also saw in Section 3 that S&P and Thomson Reuters Indices had, at least at times and for certain factors, clear multimodality. MSCI data exhibited only unimodal patterns. The effect of these differences will be examined in this section as we examine the data in a more analytical way using EDA.

In this section, the papers focus will be on cluster analysis which traditionally has played a central role in exploratory data analysis. Cluster analysis can give important clues to the structure of data sets, and therefore can suggest results and future hypotheses.

Despite being one of the most commonly used tools for unsupervised exploratory data analysis and despite its extensive literature, very little is known about the theoretical foundations of clustering methods. The general question of which methods are best, or most appropriate for a particular problem, or how significant a particular clustering is has no completely thorough grounding in mathematical theory. For example, one problem is that many clustering methods involve particular choices that need to be made at the outset. Another example is how many clusters there should be, or the value of a particular shareholding quantity. In addition, some methods depend on artifacts in the data, such as the particular order in which the elements are listed.

In this paper, we will use two cluster methods that have a very good mathematical grounding and can answer many of the questions posed above. The

first is the MCLUST procedure as developed by Fraley and Raftery (2007) and the second is from a series of papers by Carlson and others, the most recent being Carlson (2009). Fraley and Raftery's work is one of the first successful applications of a mixture of Gaussian distributions to clustering problems while Carlson *et al.*'s work suggests using *persistence* as a central criterion for assessing cluster effectiveness and significance.

Section 4.1: MCLUST

MCLUST is an R package for normal mixture modeling and model-based clustering. It provides functions for parameter estimation via the EM algorithm for normal mixture models with a variety of covariance structures. Also included in MCLUST are functions that combine model-based hierarchical clustering, EM for mixture estimation and the Bayesian Information Criterion (BIC) for clustering, density estimation and discriminate analysis.

In Fraley and Raftery (2007), the mathematical foundation of MCLUST is developed. Here a synopsis of their work will be presented. The interested reader is referred to the paper for a more complete development.

In brief, MCLUST replaces the maximum likelihood estimation (MLE) commonly used in clustering with a maximum a posteriori (MAP) estimator and arrives at a solution for the MAP using the EM algorithm. When choosing the number of components and the model parameterization, Fraley and Raftery propose a modified version of BIC, where the likelihood is evaluated via MAP instead of the MLE. Fraley and Raftery use a highly dispersed proper conjugate prior containing a small fraction of one observation's worth of information. The resulting method avoids degeneracies and singularities, but when these are not present it gives similar results to the standard method using MLE, EM and BIC.

The basic working methodology of Fraley and Raftery is:

- Specify a maximum number of components, G_{\max} , to consider, and a set of candidate parameterizations of the Gaussian model.
- Estimate parameters via EM for each parameterization and each number of components up to G_{\max} . The

conditional probabilities corresponding to a classification from model-based hierarchical clustering, or the estimated parameters for a simpler (more parsimonious) model, are good choices for initial values.

- Compute BIC for the mixture likelihood with the optimal parameters from EM for up to G_{\max} clusters.
- Select the model (parameterization / number of components) for which BIC is maximized.

Now the EM estimates can sometimes fail to converge because for many mixture models, the likelihood is not bounded and there are paths in the parameter space along which the likelihood tends to infinity.

In practice, this behavior is due to one or more singularities in the covariance estimates and arises most often for models in which the covariance is allowed to vary between components and for models with a large number of components.

Fraley and Raferty propose to avoid these problems by replacing the MLE by the maximum a posteriori (MAP) estimate from Bayesian analysis. They propose a prior distribution on the parameters that eliminates failure due to singularity, while having little effect on stable results obtainable without a prior. The Bayesian predictive density for the data is assumed to be of the form:

$$L(Y | \tau_k, \mu_k, \Sigma_k) P(\tau_k, \mu_k, \Sigma_k | \theta)$$

where L is the mixture likelihood:

$$L(Y | \tau_k, \mu_k, \Sigma_k) = \prod_{j=1}^n \sum_{i=1}^G \tau_k \phi(y_j | \mu_k, \Sigma_k)$$

$$= \prod_{j=1}^n \sum_{i=1}^G \tau_k |2\pi\Sigma_k|^{-1/2} \exp\left\{-\frac{1}{2}(y_j - \mu_k)^T \Sigma_k^{-1}(y_j - \mu_k)\right\}$$

where P is the prior distribution of parameters \bullet_k , \bullet_k and \bullet_k which includes other parameters denoted by \bullet . This technique finds a posterior mode or MAP (maximum a posteriori) estimate rather than a maximum likelihood estimate for the mixture parameters.

Fraley and Raferty continue to use BIC for model selection, but in a slightly modified form. They replace the first term on the right-hand side of the BIC equation below, two x the maximized log-likelihood, with twice the log-likelihood evaluated of the MAP.

$$BIC = 2 \log lik_m(y, \theta_k^*) - (\# params)_m \log(n)$$

As noted in the introduction of this section, both S&P and Thomson Reuters Indices show clear multimodality and maybe good candidates for the Fraley-Raferty mixture approach. MSCI data however appears to be relatively free of bimodal and multimodal data so the results of using MCLUST on this dataset will need to be interpreted carefully, because Fraley and Raferty claim that MCLUST gives results *similar to* standard modeling when MLE, EM and BIC are used.

Below is Table 1 – the parameterizations for the covariance matrix Σ_k - from Fraley and Raferty (2007).

Table 1: Parameterizations of the covariance matrix Σ_k currently available in MCLUST for hierarchical clustering (HC) and/or EM for multidimensional data. ('•' indicates availability).

identifier	Model	HC	EM	Distribution	Volume	Shape	Orientation
E		•	•	(univariate)	equal		
V		•	•	(univariate)	variable		
EII	λI	•	•	Spherical	equal	equal	NA
VII	$\lambda_k I$	•	•	Spherical	variable	equal	NA
EEI	λA	•	•	Diagonal	equal	equal	coordinate axes
VEI	$\lambda_k A$	•	•	Diagonal	variable	equal	coordinate axes
EVI	λA_k	•	•	Diagonal	equal	variable	coordinate axes
VVI	$\lambda_k A_k$	•	•	Diagonal	variable	variable	coordinate axes
EEE	$\lambda D A D^T$	•	•	Ellipsoidal	equal	equal	equal
EEV	$\lambda D_k A D_k^T$	•	•	Ellipsoidal	equal	equal	variable
VEV	$\lambda_k D_k A D_k^T$	•	•	Ellipsoidal	variable	equal	variable
VVV	$\lambda_k D_k A_k D_k^T$	•	•	Ellipsoidal	variable	variable	variable

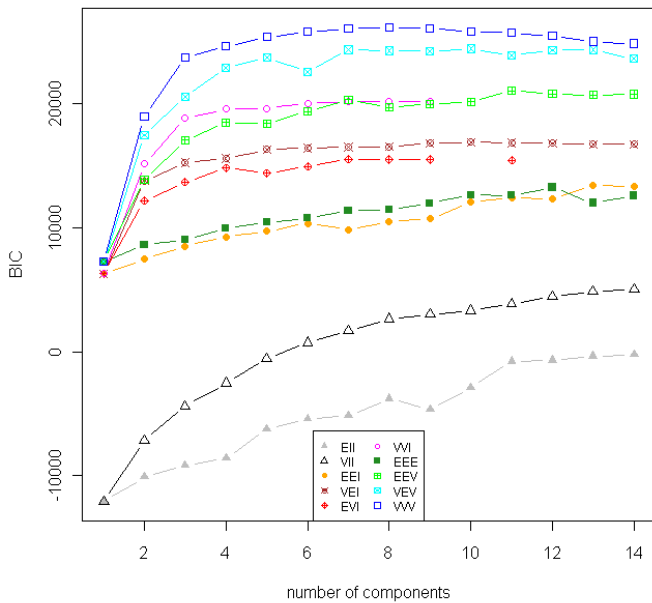
These are the model options available in MCLUST. In one dimension, there are just two models: E for equal variance and V for varying variance. In more than one dimension, the model identifiers encode geometric characteristics of the model. For example, EVI denotes a model in which the volumes of all clusters are equal (E), the shapes of the clusters may vary (V), and the orientation is the identity (I). Clusters in the EVI model have diagonal covariances with orientation parallel to the coordinate axes. The parameters associated with characteristics designated by E or V is determined by the data. A confirmation that the thick lines of both S&P and TRI are significant clusters would be if their volume and shape were E or V and their orientation V.

Section 4.1.1: S&P MCLUST Analysis

As in Section 3.1 above, we will use S&P data from 1992 as indicative of the years 1988 through 2007.

In the first test where we model the data with a Gaussian prior, we set $G_{max} = 14$ in order to examine as much of the cluster space as possible. Figure 12 has a plot of the various models tried as well as the BIC of each.

Figure 12



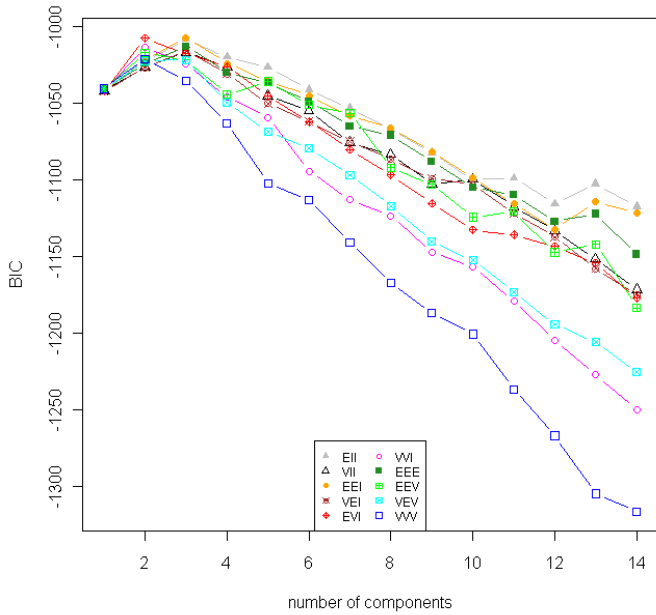
From the graph, it is clear the data is more than likely VVV or possible VEV. In either case, the data has variable volume and orientation, and quite possibly a variable shape as well. This clearly conforms to the isomap shown in Figure 3.

As to the number of clusters, the BIC values for VVV 7 – 9 are 26057.46, 26096.50 and 26163.90. All very close in value and arguments for each of them that they are the correct cluster value. The number of clusters S&P identifies using its z-scores and distance measures is at least 6 if not more as there is pure growth, pure value, a set of two that arises when the value and growth scores can both be positive and another set of two when the value and growth scores can both be negative. S&P also uses a modified decile approach (see page 6 of this paper) so the case could be made that S&P assumes 10 clusters.

Using the discriminate analysis available in MCLUST, a VVV model was specified and was trained on the odd number observations of the S&P data, with S&P data identified as pure growth, pure value, growth and value – 4 clusters. The later two classifications were determined by each stock's respective growth and value scores. The training set was then used to classify the even numbered data and overall the classification error rate was approximately 19%. Breaking this result into groups shows that pure growth was classified correctly more than 98% of the time and pure value more than 91% of the time. As these two classifications account for almost 60% of the data that means the classification error rate on what S&P calls "Stocks not in Pure Style Baskets" is approximately 54%. So MCLUST confirms the orientations and shape of the data in Figure 3 (VVV) via a mixture of distributions analysis and its discriminate analysis tool correctly classifies pure growth and pure value. However, for all practical purposes, it cannot distinguish any other stock type using the S&P variables. A further test was done by labeling the S&P stock classifications by both decile number (top 10% = 1, next 10% = 2, etc.) and using six (6) clusters (pure growth, pure value, and two additional designations for growth and value each). In each case, the discriminate analysis shows S&P's breakdown of pure growth (bottom 30%) and pure value (top 30%) is very good. Pure growth stocks have a 3% classification error rate and pure value 10%. The middle 40% in both the decile example 6-cluster example has a 66% error rate which again confirms the dispersion of attributes shown in the isomap when it comes to the stocks S&P calls "non style pure".

In Figure 13, we have a significantly different picture than we saw for S&P data in Figure 12. Here is the MSCI data from 2003 and it appears to be EEI, EVI, or EII.

Figure 13



The BIC values for EEI, EVI and EII are -1007.446, -1007.868, and -1008.628 respectively. And the number of clusters, as can be seen in Figure 13, is 2 or 3.

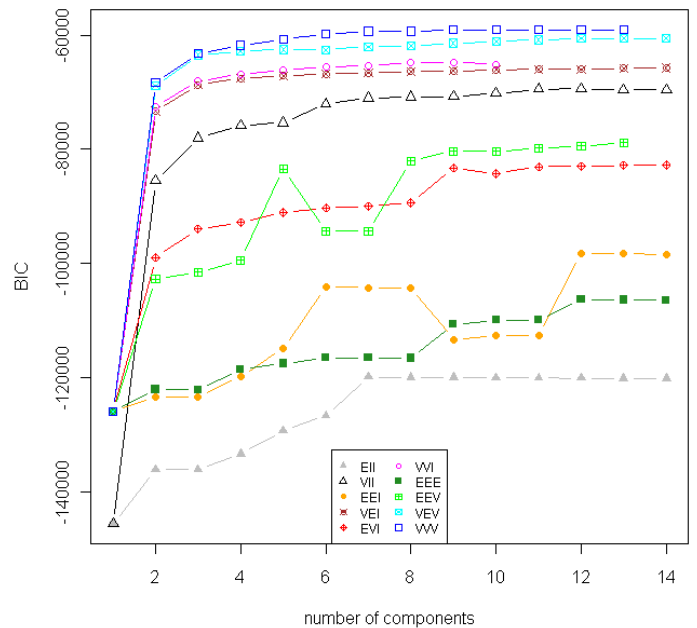
Using the discriminate analysis available in MCLUST, an EEI model was determined to be the most parsimonious and was used to train the odd number observations of the MSCI data, with MSCI data identified as pure growth, pure value, and two kinds of growth and two kinds of value. The later two classifications were determined by each stock's respective growth and value contribution scores. The training set was then used to classify the even numbered data and overall the classification error rate was approximately 35%. Breaking this result into groups shows that pure growth was classified correctly about 72% of the time and pure value more than 77% of the time. As these two classifications account for almost 55% of the data that means the classification error rate on what MSCI calls "both growth and value and non-growth non-value" stocks is approximately 61%.

So MCLUST gives us no significant insight into the shape of the MSCI data seen in Figures 7 and 8 while the discriminate analysis tool questions MSCI's pure growth and pure value classification

methodology as well as the other classifications. This is an unfortunate result but possibly not an unexpected one given the topological space the MSCI data generates – a widely scattered set of points that appears to not have any clear clusters. It also suggests that the lack of clusters and the nonlinearity noted in both graphs in Figure 5 may be working against MSCI's linear methodology. As mentioned before, the MSCI data is the shortest dataset in this paper so the MCLUST results must be interpreted tentatively given the relative paucity of data.

In Figure 14, we see that the isomap of the 1984 Thomson Reuters Indices data is confirmed by MCLUST's mixture of distributions:

Figure 14



The VVW or VEV configuration is variable in volume, variable or equal in shape, and variable in orientation.

Using the discriminate analysis available in MCLUST, a VVW model was determined to be the most parsimonious. VVW was used to train the odd number observations of the Thomson Reuters Indices data, with Thomson Reuters Indices data identified as growth, value and core. The training set was then used to classify the even numbered data and overall

the classification error rate was approximately 3%. Breaking this result into groups shows that core, growth and value all had approximately a 3% error rate. So the Thomson Reuters Indices model, so far, shows the best classification results, especially for that middle 40% of stocks that Thomson Reuters Indices is labeling core.

Section 4.2: Clustering and Algebraic Topology

In a recent article Carlsson (2009) discusses how geometry and topology can be applied to analyze various kinds of data. He notes that geometry and topology are very natural tools to apply to point cloud data since geometry can be regarded as the study of distance functions and what one often works with in large point cloud data structures are distance functions.

Carlsson also claims that geometry and topology's joint mathematical formalism can be adapted to build tools to study point clouds. Point clouds can be thought of as finite samples taken from a geometric object, perhaps with noise. Here are some of the key topics according to Carlsson that geometric methods can cover:

Qualitative information is needed: One important goal of data analysis is to allow the user to obtain *knowledge* about the data, i.e. to understand how it is organized on a large scale. For example, if we imagine that we are looking at a data set constructed somehow from diabetes patients, it would be important to develop the understanding that there are two types of the disease, namely the juvenile and adult onset forms. Once that is established, one of course wants to develop quantitative methods for distinguishing them, but the first insight about the distinct forms of the disease is key.

Metrics are not theoretically justified: In physics, the phenomena studied often support clean explanatory theories which tell one exactly what metric to use. In biological [or economics] problems, on the other hand, this is much less clear. In the biological context, notions of distance are constructed using some intuitively attractive measures of similarity but it is far from clear how

much significance to attach to the actual distances, particularly at large scales.

Coordinates are not natural: Although we often receive data in the form of vectors of real numbers, it is frequently the case that the coordinates, like the metrics mentioned above, are not natural in any sense, and that therefore we should not restrict ourselves to studying properties of the data which depend on any particular choice of coordinates. Note that the variation of choices of coordinates does not require that the coordinate changes be rigid motions of Euclidean space. It is often a tacit assumption in the study of data that the coordinates carry intrinsic meaning, but this assumption is often unjustified.

Summaries are more valuable than individual parameter choices:

One method of clustering a point cloud is the so-called *single linkage clustering*, in which a graph is constructed whose vertex set is the set of points in the cloud where two such points are connected by an edge if their distance is $\leq \bullet$, where \bullet is a parameter. Some work in clustering theory has been done in trying to determine the optimal choice of \bullet but it is now well understood that it is much more informative to maintain the entire *dendrogram* of the set which provides a summary of the behavior of clustering under all possible values of the parameter \bullet at once. It is therefore productive to develop other mechanisms in which the behavior of invariants or construction under a change of parameters can be effectively summarized.

Carlsson goes on to state that informally clustering refers to the process of partitioning a set of data into a number of parts or clusters which are distinguishable from each other. In the context of finite metric spaces, this means that points within the clusters are nearer to each other than they are to points in different clusters. Clustering then can be thought of as the statistical counterpart to the geometric construction of the path-connected components of a space, which is the fundamental building block of algebraic topology. There are many schemes which construct clusterings based on metric information, such as single, average, and complete linkage clustering, k -means clustering, spectral clustering, etc. And although clustering is a very important part of data analysis, the ways in which it is formulated and implemented is fraught with

ambiguities. In particular, the arbitrariness of various threshold choices and lack of robustness are typical of the difficulties analysts face. Much of Carlsson's and others recent research efforts has been focused on these problems and Carlsson's work in particular has shown that functoriality⁵ can provide the right general mathematical framework for addressing them. For example, one can construct data sets which have been thresholded at two different values, and the behavior of clusters under the inclusion of the set with tighter threshold into the one with the looser threshold is informative about what is happening in the data set.

Fraley and Raftery's MCLUST routine addresses some of these issues such as the arbitrariness of various threshold choices and lack of robustness via their use of multiple modeling of the cluster space (EEE, VVI, VEV, etc.) and determining the better model via EM and BIC. But they do not address an important point Carlsson raised above, that "notions of distance are constructed using some intuitively attractive measures of similarity ... but it is far from clear how much significance to attach to the actual distances, particularly at large scales." This is a point the author emphasized in Section 3 as the index providers' assumptions of the euclidean or manhattan metric do not seem to be justified by their respective point cloud topologies.

In Carlsson and Memoli (2008), an argument is made in favor of using persistence as a way of encoding multiscale or multiresolution into cluster analysis. Their definition of persistence is given below. A full discussion of persistence is in Carlsson (2009).

Let $P(X)$ denote the partitions of the finite set X .

⁵ Carlsson (2009) define functoriality as, "The relationships which are useful involve continuous maps between the different geometric objects, and therefore become a manifestation of the notion of *functoriality*, i.e., the notion that invariants should be related not just to objects being studied, but also to the maps between these objects. Functoriality is central in algebraic topology in that the functoriality of homological invariants is what permits one to compute them from local information, and that functoriality is at the heart of most of the interesting applications within mathematics. Moreover, it is understood that most of the information about topological spaces can be obtained through diagrams of discrete sets, via a process of simplicial approximation."

Definition: A persistent set is a pair (X, \bullet) where X is a finite set and \bullet is a function from the non-negative real-line $[0, +\infty)$ to $P(X)$ so that the following properties hold

1. If $r \leq s$ then $\bullet(r)$ defines $\bullet(s)$
2. For any r , there is an $\epsilon > 0$ so that $\bullet(r') = \bullet(r)$ for all $r' \in [r, r + \epsilon]$

If in addition there exists a t s.t. $t > 0$, $\bullet(t)$ consists of a single block partition for all $r \geq t$, then we say (X, \bullet) is a dendrogram.

Carlsson and Memoli go onto explain:

Since there are only a finite number of partitions of X , a persistent set Q gives a partition of \mathbb{R}^+ into a finite collection I of intervals of the form $[r, r')$ together with one interval of the form $[r, +\infty)$. For each such interval, every number in the interval corresponds to the same partition of X .

We claim that knowledge of these intervals is a key piece of information about the persistent sets arising in cluster analysis. The reason is that long intervals in I correspond to large ranges of values of the scale parameter in which the associated cluster decomposition doesn't change. One would then regard the partition into clusters corresponding to that interval as likely to represent significant structure present at the given range of scales. If there is only one long interval (aside from the infinite interval of the form $[r, +\infty)$) in I , then one is led to believe that there is only one interesting range of scales, with a unique decomposition into clusters. However, if there is more than one long interval, then it suggests that the object has significant multiscale behavior. Of course, the determination of what is "long" and what is "short" will be problem dependent, but choosing thresholds for the length of the intervals will give definite ranges of scales. As for the computability, the persistent sets associated to a finite metric space can be readily computed using (conveniently modified) hierarchical clustering techniques or the methods of persistent homology.

In R, a routine called PVCLUST has the "convenient modification" Carlsson and Memoli suggest above. With PVCLUST, the analyst can choose the relative sizes of the bootstrap replicates as well as a method of clustering and a distance metric. Two significance tests are available to determine which clusters have significant multiscale behavior,

Suzuki and Shimodaira, the developers of PVCLUST, state in a paper from 2004 that in general, the result of hierarchical clustering for p individuals contains $p - 1$ clusters. However, it is not clear how strong a cluster is supported by the data.

PVCLUST measures the accuracy of these clusters as p -values, which ranges from 0 to 1. If the p -value of a cluster is less than α , the cluster is rejected at the α level of significance. Multiscale bootstrap resampling is a method which calculates p -values of hypotheses by resampling of data.

Note that their p -value calculated by multiscale bootstrap resampling is an approximation. But it is less biased than b -value bootstrap probability which is also an approximation of p -value computed by bootstrap resampling.

Figure 15 is from the 2004 paper: a dendrogram for lung tumor data.

Figure 15

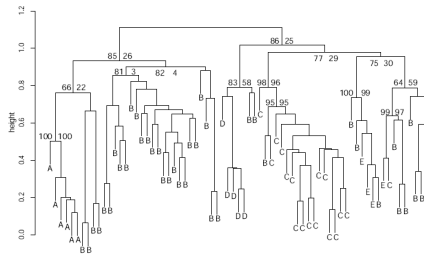
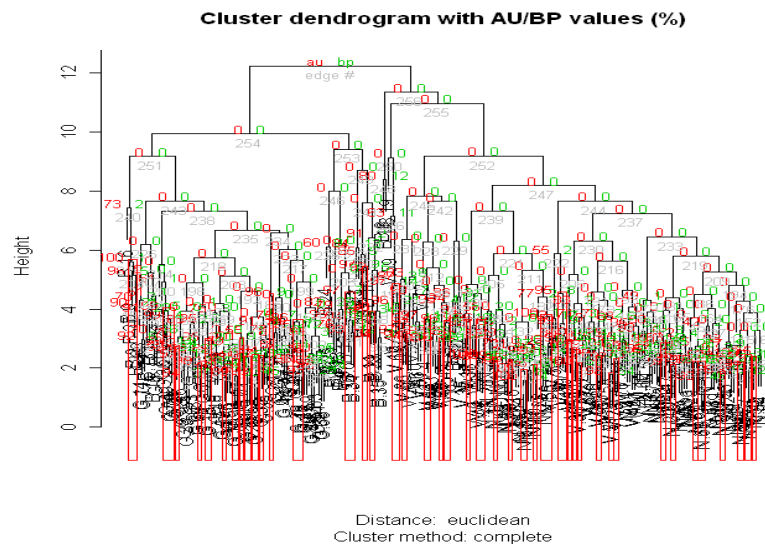


Figure 1: Hierarchical cluster analysis of 73 lung tumors. Values at branches are p -values (left), bootstrap probabilities (right) in percentage. Labels at leaves (A, B, C, D) are classification by specialists.

The numbers on top of the branches (edges) are p and b values with p to the left and b to the right. Those edges with $p \geq 95\%$ are taken as significant clusters.

In Figure 16 a dendrogram using MSCI data shows a significant number of clusters but this comes with a number caveats. Please bear in mind that labels will be black smears in the dendrograms that follow as the number of stocks under examination exceed 1000.

Figure 16



The clusters bounded in red are the statistically significant clusters, i.e. those clusters that satisfy the Carlsson and Memoli claims of persistence. Note that the distance metric is euclidean and the method is complete hierarchical which uses the furthest neighbor rule, i.e., the largest dissimilarity between and objects in cluster A and objects in cluster B determines the members and numbers of clusters.

There are 43 clusters that are significant but the number of stocks that are contained in those clusters only amount to 41% of all the stocks to be classified. 69% of the 43 clusters are composed of a single kind of stock, e.g., all growth or all value, but 31% are a mix of stock types.

An analysis of the 59% of stocks not classified finds no significant difference or bias towards growth, value or what lies in between. Each fails not to be classified at about the same level in term of the number of unclassified stocks to total stocks.

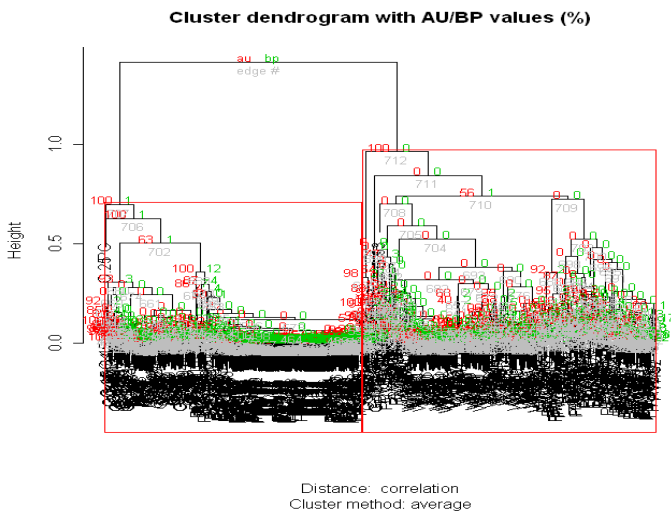
Different methods of hierarchical clustering were tried as were different distance metrics. Material difference were found when using other techniques or metrics with those differences being more often that not more stocks were being classified but with biases appearing in the not classified group, e.g., growth stocks dominating the non-classifiable in certain techniques. The best method in terms of numbers of stocks classified was 57% and that was with the ward cluster technique. This makes

geometrical sense as the ward technique looks for spherical clusters and spherical clusters could certainly capture the some of the topology of the MSCI's 3D isomap. The problem with the ward result is that value stocks do ok but other stock types in the majority of cases either get misclassified or not classified at all.

It is the author's hypothesis that the lack of linearity of factors MSCI uses that is the main reason for the poor performance of the MSCI classification methodology when either a topology-preserving or multiple modeling method is used. It appears that the MSCI methodology does not do justice to the MSCI point cloud. Again, the reader must keep in mind that only nine years of MSCI data are available so drawing conclusions on such a small sample must be made with care.

In Figure 17, we have the dendrogram for S&P data:

Figure 17



Here we have two clusters. The one on the left is made up of entirely of growth stocks and the one on the right is 95% value stocks. And the clusters include all the stocks that need to be classified so no stocks using the S&P methodology goes unclassified, unlike what happened with MSCI.

So a very good split of stocks and very much what would be expected given the S&P isomap in Figure 3. Again other techniques and metrics, in particular single linkage which is close to the *knn*

methodology used to create the isomaps were tried but complete linkage and the euclidean metric used to generate Figure 3 was clearly the best. The conclusion here is that S&P's classification methodology does a very good job of separating growth and value stocks. Their claim however to be able to identify pure growth, pure value and a third category called "non-pure style basket" seems to be in jeopardy given the results of this and the MCLUST tests.

The Thomson Reuters data could not be analyzed via PVCLUST as no robust methodologies are available. The author used BOOT, the R bootstrap package as a wrapper for the R package CLUST. CLUST does allow the use of robust techniques in its hierarchical clustering routines. Appendix A has the cluster breakouts for Thomson Reuters Indices and similar to S&P, the clusters are very clean and all stocks are classified. The difference Thomson Reuters Indices has with S&P is that a third cluster does appear that is made-up entirely of what Thomson Reuters Indices calls core.

The misclassification percentages of Thomson Reuters Indices are higher than S&P - 10% for value and 7% - for growth but these are still good results.

Section 4.3: Testing for the Value Effect

In the introduction to this paper, the author made mention of the value effect – periods of time when value stocks as a group outperform growth stocks. In this brief section, how each index providers' growth and value indices measure up against the accepted benchmark of the value effect – the Fama-French benchmarks – will be discussed.

Both S&P and MSCI have a very mixed performance versus Fama-French. Some years when the value effect is in effect, their value indices do better than their growth indices while in other years the opposite happens. There appears to be no pattern to their performance against the Fama-French benchmarks so investors, portfolio managers and other users of S&P and MSCI growth and value indices should keep in mind this erratic performance and keep a separate eye on the Fama-French benchmarks.

The Thomson Reuters Indices in almost every year mimic the Fama-French results (which should not be

surprising as Thomson Reuters' methodology is very similar to Fama-French). However, the Thomson Reuters Indices *always* underperform Fama-French. So though Thomson Reuters gets the direction right more often than not, the value lags behind the benchmark.

In the conclusion of this paper, the author will discuss an interesting statistical anomaly (or artifact) that raises some questions about the value effect.

Section 5: Conclusions

This paper has shown the importance of understanding the topology and geometry of a datasets' point clouds. With this information, the analyst can make informed hypotheses concerning what techniques to possibly use, what the best metrics maybe, and in general be better informed when interpreting the results of tests done on the data.

The paper also introduced an important advance in algebraic topology that could have a significant impact on statistical clustering, pattern recognition and the like in the future. Carlsson and Carlsson-Memoli give good mathematical grounding to the work of Suzuki and Shimodaira and the work found in such texts as *The Handbook of Genome Research*⁶ where hierarchical multiresolution bootstrap resampling is used to study biological test results, in particular applications of same in bioinformatics and genomics.

In terms of the datasets in this paper, both S&P and Thomson Reuters Indices demonstrated early on that their point clouds could support a linear separation of growth and value stocks. MSCI's point cloud which could be described in a variety of ways but is clearly nonlinear, raises important questions about MSCI's assumptions concerning the use of its linear methodology.

The tests also showed that classifying that central core of stocks is very difficult to do. Thomson Reuters Indices does not attempt to do so, so the results of their clustering tests indicate that not doing so is a good idea. S&P, which does try to classify the central core, does not do so at all

successfully. As a matter of fact, the multiscale bootstrap results indicate that only growth and value can be identified given S&P's factors and methodology. This raises an interesting question that this paper cannot answer – is there a group of factors and techniques that can distinguish what S&P (and MSCI) are looking for? Or is it a mirage? The answer to this question will be left to future research.

Outliers and missing variables were briefly addressed in this paper and in most cases they were found not to significantly affect the classification of stocks. However, it should be borne in mind that Thomson Reuters Indices uses a robust methodology and MSCI winsorizes its data. And it was noted that the non-linearity in S&P's manifolds appears to be caused by extreme values and outliers. It is the author's suggestion that S&P consider the using *A-estimators* to start as eliminators of outliers as they are very good estimators when multimodality is present.

Thomson Reuters Indices, like all the others, could benefit from a better understanding of the geometry and topology of its point clouds. That they use a robust technique to classify stocks is clearly to their benefit, given the outliers seen in their isomaps. But changes in the topology of the fundamental factors, especially a move away from linear clusters, would severely damage their methodology, much as it would S&P's. A multi-dimensional tool is needed by Thomson Reuters Indices (and S&P) in order to better understand and better separate growth and value stocks.

Finally, MSCI has the more pressing need to either identify a set of fundamental factors that are amenable to linear classification methodologies or move in the direction suggested for Thomson Reuters Indices and S&P – develop a multi-dimensional classification tool. A possible starting point for MSCI would be Carlsson [2009], especially the section about Cech and Vietoris-Rips complexes. It is clear from the 3D isomap that if the visible edges are truly part of a Delaunay complex giving the bounds of one or more convex hulls then Delaunay triangulation could open up some interesting doors. Another avenue to examine is to see if the surface is one of submanifolds of Euclidean space, such as the one noted for Fletcher Powell, one could construct the restricted Delaunay triangulations.

⁶ *The Handbook of Genome Research*, Vol. 2, Chapter 17, Wiley-VCH, 2005.

Finally, this paper is the first in what is hoped to be a two or three-part series about style properties in stocks. The next paper will be about the use of computational agent modeling to study cross-sectional stock returns, in particular the value effect. As mentioned earlier, there is no agreement in the economics community about the cause of the value effect though this author does have a hypothesis – the value effect could be a growth effect. In other words, the book-to-value deciles that Fama-French, Lakonishok-Chan and others talk about comes about because the lower two deciles (the most “growthy” stocks) are *always* statistically different from the deciles above them, i.e., they clearly come from a different distribution. That cannot be said with *any consistency* about deciles 1 thru 7. So the value effect may not be about irrationality or risk. But that is another question left to future research.

Acknowledgements

The author would like to thank Dr. Mark Labovitz for his initial insights on the problem of fundamental data multimodality. Without that initial conversation with Mark, this paper would not have appeared in its current form. The author would also like to thank Professor Gunnar Carlsson and the members of the Applied Topology group at Stanford. Their comments and insights provided valuable improvements to this paper. Any errors in this paper are the author’s.

REFERENCES

- Back, T., *Evolutionary Algorithms in Theory and Practice*, Oxford, 1996.
- Carlsson, G., "Topology and Data," *Bulletin of the American Mathematical Society*, 46, 255-308, 2009.
- Carlsson, G. and Memoli, F., "Persistent Clustering and a Theorem of J. Kleinberg," arXiv:0808.2241v1 [stat.ML] 16 Aug 2008.
- Chan, L., Karceski, J. and Lakonishok, J., "New Paradigm or Same Old Hype in Equity Investing," *Financial Analyst Journal*, 56, 23-36, 2000.
- Chan, L. and Lakonishok, J., "Value and Growth Investing: Review and Update," *Financial Analyst Journal*, 60: 71-86, 2004.
- Fama, E. F. and French, K. R., "The Cross-Section of Expected Stock Returns" *Journal of Finance* 47, 427-465, 1992.
- Fama, E. F. and French, K. R. (1996). "Multifactor Explanations of Asset Pricing Anomalies," *Journal of Finance* 51, 55-84, 1996.
- Fraley, C. and Raftery, A., "MCLUST Version 3 for R: Normal Mixture Modeling and Model-Based Clustering," Technical Report 504, University of Washington, Department of Statistics, 2007.
- Marchette, D., *Random Graphs and Statistical Pattern Recognition*, Wiley, 2004.
- Roll, R and Ross, S., "An empirical investigation of the arbitrage pricing theory". *Journal of Finance*, 35, 1073-1103, 1980.
- Stutzer, M., "Fund Managers may cause their Benchmarks to be Priced Risks," *Journal of Investment Management*, 1, 1-13, 2003.
- Suzuki, R. and Shimodaira, H., "An application of multiscale bootstrap resampling to hierarchical clustering of microarray data: How accurate are these clusters?" <http://www.is.titech.ac.jp/~shimo/pub/GIW2004/GIW04P034.pdf>

APPENDIX

Thomson Reuters Indices Data from Multiscale Bootstrap Resampling

In the clusters below "V" refers to value, "C" to core and "G" to growth

Cluster 1:

V	V.1	V.2	V.3	V.4	V.5	V.6	V.7	V.8
V.9	V.10	V.11	V.12	V.13	V.14	V.15	V.16	V.17
V.18	V.19	V.20	V.21	V.22	V.23	V.24	V.25	V.26
V.27	V.28	V.29	V.30	V.31	V.32	V.33	V.34	V.35
V.36	V.37	V.38	V.39	V.40	V.41	V.42	V.43	V.44
V.45	V.46	V.47	V.48	V.49	V.50	V.51	V.52	V.53
V.54	V.55	V.56	V.57	V.58	V.59	V.60	V.61	V.62
V.63	V.64	V.65	V.66	V.67	V.68	V.69	V.70	V.71
V.72	V.73	V.74	V.75	V.76	V.77	V.78	V.79	V.80
V.81	V.82	V.83	V.84	V.85	V.86	V.87	V.88	V.89
V.90	V.91	V.92	V.93	V.94	V.95	V.96	V.97	V.98
V.99	V.100	V.101	V.102	V.103	V.104	V.105	V.106	V.107
V.108	V.109	V.110	V.111	V.112	V.113	V.114	V.115	V.116
V.117	V.118	V.119	V.120	V.121	V.122	V.123	V.124	V.125
V.126	V.127	V.128	V.129	V.130	V.131	V.132	V.133	V.134
V.135	V.136	V.137	V.138	V.139	V.140	V.141	V.142	V.143
V.144	V.145	V.146	V.147	V.148	V.149	V.150	V.151	V.152
V.153	V.154	V.155	V.156	V.157	V.158	V.159	V.160	V.161
V.162	V.163	V.164	V.165	V.166	V.167	V.168	V.169	V.170
V.171	V.172	V.173	V.174	V.175	V.176	V.177	V.178	V.179
V.180	V.181	V.182	V.183	V.184	V.185	V.186	V.187	V.188
V.189	V.190	V.191	V.192	V.193	V.194	V.195	V.196	V.197
V.198	V.199	V.200	V.201	V.202	V.203	V.204	V.205	V.206
V.207	V.208	V.209	V.210	V.211	V.212	V.213	V.214	V.215
V.216	V.217	V.218	V.219	V.220	V.221	V.222	V.223	V.224
V.225	V.226	V.227	V.228	V.229	V.230	V.231	V.232	V.233
V.234	V.235	V.236	V.237	V.238	V.239	V.240	V.241	V.242
V.243	V.244	V.245	V.246	V.247	V.248	V.249	V.250	V.251
V.252	V.253	V.254	V.255	V.256	V.257	V.258	V.259	V.260
V.261	V.262	V.263	V.264	V.265	V.266	V.267	V.268	V.269
V.270	V.272	V.273	V.274	V.275	V.276	V.277	V.278	V.279
V.280	V.281	V.282	C	C.1	C.2	C.3	C.4	C.5
C.6	C.7	C.8	C.9	C.10	C.11	C.12	C.13	C.14
C.15	C.16	C.17	C.18	C.19	C.20	C.21	C.22	C.23
C.24	C.25	C.26	C.27	C.28	C.29			

Cluster 2:

						C.30	C.31	C.32
C.33	C.34	C.35	C.36	C.37	C.38	C.39	C.40	C.41
C.42	C.43	C.44	C.45	C.46	C.47	C.48	C.49	C.50
C.51	C.52	C.53	C.54	C.55	C.56	C.57	C.58	C.59
C.60	C.61	C.62	C.63	C.64	C.65	C.66	C.67	C.68
C.69	C.70	C.71	C.72	C.73	C.74	C.75	C.76	C.77

C.78	C.79	C.80	C.81	C.82	C.83	C.84	C.85	C.86
C.87	C.88	C.89	C.90	C.91	C.92	C.93	C.94	C.95
C.96	C.97	C.98	C.99	C.100	C.101	C.102	C.103	C.104
C.105	C.106	C.107	C.108	C.109	C.110	C.111	C.112	C.113
C.114	C.115	C.116	C.117	C.118	C.119	C.120	C.121	C.122
C.123	C.124	C.125	C.126	C.127	C.128	C.129	C.130	C.131
C.132	C.133	C.134	C.135	C.136	C.137	C.138	C.139	C.140
C.141	C.142	C.143	C.144	C.145	C.146	C.147	C.148	C.149
C.150	C.151	C.152	C.153	C.154	C.155	C.156	C.157	C.158
C.159	C.160	C.161	C.162	C.163	C.164	C.165	C.166	C.167
C.168	C.169	C.170	C.171	C.172	C.173	C.174	C.175	C.176
C.177	C.178	C.179	C.180	C.181	C.182	C.183	C.184	C.185
C.186	C.187	C.188	C.189	C.190	C.191	C.192	C.193	C.194
C.195	C.196	C.197	C.198	C.199	C.200	C.201	C.202	C.203
C.204	C.205	C.206	C.207	C.208	C.209	C.210	C.211	C.212
C.213	C.214	C.215	C.216	C.217	C.218	C.219	C.220	C.221
C.222	C.223	C.224	C.225	C.226	C.227	C.228	C.229	C.230
C.231	C.232	C.233	C.234	C.235	C.236	C.237	C.238	C.239
C.240	C.241	C.242	C.243	C.244	C.245	C.246	C.247	C.248
C.249	C.250	C.251	C.252	C.253	C.254	C.255	C.256	C.257
C.258	C.259	C.260	C.261	C.262	C.263	C.264	C.265	C.266
C.267	C.268	C.269	C.270	C.271	C.272	C.273	C.274	C.275

Cluster 3:

C.276	C.277	C.278	C.279	C.280	C.281	C.282	C.283	C.284
C.285	C.286	C.287	G	G.1	G.2	G.3	G.4	G.5
G.6	G.7	G.8	G.9	G.10	G.11	G.12	G.13	G.14
G.15	G.16	G.17	G.18	G.19	G.20	G.21	G.22	G.23
G.24	G.25	G.26	G.27	G.28	G.29	G.30	G.31	G.32
G.33	G.34	G.35	G.36	G.37	G.38	G.39	G.40	G.41
G.42	G.43	G.44	G.45	G.46	G.47	G.48	G.49	G.50
G.51	G.52	G.53	G.54	G.55	G.56	G.57	G.58	G.59
G.60	G.61	G.62	G.63	G.64	G.65	G.66	G.67	G.68
G.69	G.70	G.71	G.72	G.73	G.74	G.75	G.76	G.77
G.78	G.79	G.80	G.81	G.82	G.83	G.84	G.85	G.86
G.87	G.88	G.89	G.90	G.91	G.92	G.93	G.94	G.95
G.96	G.97	G.98	G.99	G.100	G.101	G.102	G.103	G.104
G.105	G.106	G.107	G.108	G.109	G.110	G.111	G.112	G.113
G.114	G.115	G.116	G.117	G.118	G.119	G.120	G.121	G.122
G.123	G.124	G.125	G.126	G.127	G.128	G.129	G.130	G.131
G.132	G.133	G.134	G.135	G.136	G.137	G.138	G.139	G.140
G.141	G.142	G.143	G.144	G.145	G.146	G.147	G.148	