Does a Cluster Based Factor Model Perform Better Than an Industry Based Factor Model? An Investigation of European Equities

Kumar Gautam

Master of Science in Finance (2011-12)

EDHEC Business School, Nice

Abstract

Past research that analyse the influence of country and industry factors in determining international stock returns fail to reach a consensus on the relative importance of country and industry effects. This leaves an international equity portfolio manager in dilemma whether to allocate money based on country considerations or industry considerations. Towards this end, I try to analyze an alternative that has the potential to extract maximum benefit from the structure of international stock returns. I cluster stocks which behave similarly, and then use them as inputs to a factor model for international stock returns. I compare a factor model based on clusters with a factor model based on industry classifications. I conclude that cluster based factor model perform better than an industry based factor model.

I express my sincere gratitude to Professor Abraham Lioui for providing valuable guidance to complete this research paper. I am particularly grateful to Professor Sergio Focardi who helped me with his inputs at various stages of this research paper. I sincerely thank EDHEC Business School for providing excellent learning environment. Obviously, any errors and unjustified conjectures in this research paper are my own.

1. Introduction

An international equity portfolio manager often faces two competing choices in allocating funds. If the portfolio manager thinks that country factors are more important in determining international stock returns, he may either allocate money to country based indices or take exposure to selected stocks in those countries. On the contrary, if the manager thinks industry factors are more important in determining international stock returns, he may either allocate money to industry indices or take exposure in selected stocks in those industries. In either case, the aim is to extract the maximum benefit from the structure of international stock returns.

Towards this end, I try to analyze an alternative that has the potential to extract maximum benefit from the structure of international stock returns. This paper focuses on European equity market. I cluster stocks as per their historical behavior, and I treat these clusters as factors. I assume that realized return on a stock can be expressed as a function of a factor common to all stocks, an effect of the cluster to which the stock belongs to and an idiosyncratic disturbance.

This model has been adopted from Heston and Rouwenhorst (1995). In this seminal paper, they have expressed a stock return as a function of a factor common to return on all the stocks, an effect of the industry to which the stock belongs to, an effect of the country to which the stock belongs to and an idiosyncratic disturbance. They used cross-sectional regression to estimate country and industry effects, and have assigned each stock a beta (dummies) of 1/0 depending on whether the stock belongs to a particular country or industry. Since, then studies [for example see Cavaglia, Brightman and Aked (2000), Hamelink, Harasty and Hillion (2001)] that try to understand and compare the importance of country and industry factors, have used a similar model.

Treating clusters of stocks as factors can be interpreted in the same way as other classical asset pricing theories, although it requires a formal theoretical structure. Single index model in Sharpe (1970) allows to express expected return on an asset with respect to a broad market index and company specific risk, and multifactor model in Ross (1976) allows to break return on an asset into multiple systematic sources and an idiosyncratic component.

Cluster based factor models can be interpreted in a similar manner. Clustering attempts to group stocks in such a way that behavior of stocks *within a cluster* is same but the behavior of a cluster with another cluster is different. It can, therefore, be conjectured that stocks within a cluster react similarly to different systematic factors, and these systematic factors are specific to a particular cluster. This underlines the logic behind expressing return on a stock as return associated with the cluster to which that stock belongs to, a common factor shared by all the clusters and a company specific factor.

An obvious drawback of this model is that there may be overlapping systematic factors. If the distance between clusters is large, it can be assumed that the clusters are impacted by factors specific to them. However, if the distance between clusters is small, it can be said that there are overlapping systematic factors. It is, therefore, important to find whether such clusters (with small *within cluster* distance and large *in between cluster* distance) exist in the structure of stock returns.

Many studies justify the logic behind clustering stocks and using cluster effects to express stock returns. Farrell, Jr. (1974) used clustering to identify clusters with small cluster within cluster distance and large in between cluster distance. The study aims to address the issue of collinearity among indxes in multi-index models. Arnott (1980) exploits correlation between stocks to cluster stocks and express stock return as a linear combination of cluster effects and stock's beta corresponding to that cluster. It also highlights that cluster betas change over time, but they are more stable than the market beta in the single index model. Mantegna (1999) spotted hierarchical structure in the S&P 500 stocks using the date from July 1989 to October 1995. Mantegna mentions that stocks are affected by factors specific to its corresponding cluster and these clusters can be used to theoretically explain the behavior of stock returns.

It is imperative to justify why this paper compares a cluster based factor model with an industry based factor model and why not it simply analyse the effect of clustering on factor models in general. This paper is not a theoretical exercise to explain the behavior of international stock returns. Instead, the idea is to add to the practice of multifactor risk analysis for an international portfolio manager.

Already decades of academic research have went into analyzing the significance of country and industry factors in determining international stock returns, but no consensus has been reached yet. Early studies such as Lassard (1974), Solinik (1974), Beckers, Connor and Curds (1996), Grinold, Rudd and Stefek (1989) and Heston and Rouwenhorst (1994) conclude that national factors are more dominant that industry factors. However, the more recent studies such as Hamelink, Harasty and Hillion (2001) and Cavaglia, Brightman and Aked (2000) have seem to reach a consensus that industry factors are becoming more dominant in determining international stock returns. Also, as these studies focus on different markets and time horizon, it is not clear which factor actually dominate international stock returns.

To cut through the confusion, I use a completely different method to identify dominant factors for international stock returns. To understand if this method of constructing factor model has an added value, I compare it to an industry based factor model, and this justifies the selection of the topic for this research paper.

I analyse weekly return data from the January 2007 to December 2011 for the European equities from the 18 countries. I use cross-sectional regression, based on the model proposed in Heston and Rouwenhorst (1995), for each week to estimate time series of pure industry and cluster effects. Next, I check the significance of these factor returns and how well the model fits the data. Further, I analyse potential profitability of the two different strategies: one that is based purely on industry factors and the other that is based purely on cluster factors. I also compare the diversification potential for these two strategies.

I find that both industry and cluster effects are dominant in European stock market, but cluster based returns are more significant than industry based returns. Also, a strategy based on clustering techniques has higher profitability and higher diversification potential. I conclude that clustering approach to multifactor risk models adds a new dimension in analyzing and profiting from structures present in international stock returns.

2. Literature Review

Clustering is a technique to extract information from noisy data. Its aim is to classify objects into groups (called clusters) in a manner that minimize distance between objects within a cluster and maximize distance between the clusters. It has application in areas as wide as medicines, genetics, marketing and sociology.

In finance, clustering is widely used to extract information from noisy time series data. At the outset, it is important to distinguish two different ways in which clustering has been used in finance. First, clustering is used as a pure data mining tool, with little economic intuition behind the groups obtained as a result of the clustering exercise. Second, clustering can be used as an econometric tool in a manner so that economic intuition justifies the groups that are formed as a result of the clustering exercise.

This paper uses clusters as inputs to model factors for international stock returns. Even before the statistical significance of these clusters is examined, it is important to know whether time series of stock returns can be grouped into clusters that have economic interpretation. If not, the whole exercise of using clusters as factors will be rendered as pure data mining exercise.

Thus, this section of the research paper first reviews previous research that establishes the existence of economically relevant clusters in stock returns. Also, I review literature that either uses clusters to express stock returns or have suggested that economically meaningful clusters can be used to express stock returns. Next, I briefly review the concept of factor models to put things in perspective. Since I use a cluster based factor model that has a structure similar to industry based factor model, and also since I compare cluster based factor model with industry based factor model to highlight its relevance, I review literature related to industry and country based factor models in detail.

2.a Clustering time series of stock returns

Past studies apply clustering techniques for stock return data to achieve different objectives, but the underlying theme that emerges is that economically meaningful clusters can be extracted using the historical behavior of stock returns.

Farrell, Jr. (1974) clustered stocks and analysed if cluster effects can explain cross-sectional variation in stock returns. The study analysed monthly return data of 100 stocks from 1961 to 1969. Before forming clusters, the impact of market factor was isolated from the stock returns. Each month, the cross section of stock returns was regressed over market return and time series of residual for each stocks were obtained. The correlation between the residuals were used to form four different clusters (Farrell initially hypothesized three clusters but an additional fourth cluster was obtained). These groups, in addition to market index, were shown to explain 45 % variation in stock return as compared to 31 % explained by market index.

The study also highlighted that covariance matrix obtained using multi-index model is found to produce less efficient portfolios. It is due to the fact that indices used in multi-index model exhibited collinearity. Farrell’s aim of clustering was also to form clusters, which are non-collinear in nature.

Arnott (1980) emphasized the need to identify systematic risks other than the market risk in single index model, so as to understand the co-movement of stock with groups other than the market index. The author forms five clusters in addition to market factor, and used them in the multifactor model. The five factors combined have an additional 9.2 % explanatory power as compared to 29.5 % explanatory power of market index. This suggests that although market beta and company specific factor have major explanatory power, the clusters highlight new dimensions of risk, which have explanatory power large enough that can't be ignored. The author particularly highlighted the application of cluster based multifactor model in optimization and performance management. This could lead to better understanding of sources risk and return.

Mantegna (1999) applies clustering technique to confirm the presence of hierarchical structure in stock return data. The sample was stocks in Dow Jones Industrial Average index and S&P 500 index between the period July 1989 and October 1995. Unlike previous studies that formed multifactor models using clusters, Mantegna focused on applying clustering techniques to identify economically meaningful clusters. He used correlation as the measure of distance, but manipulated the correlation coefficient so that it satisfies the basic property of distance that it can’t be negative (note that correlation can be negative whereas distance can't be negative).

Mantegna mentions that clusters are exposed to risks specific to them (although there might be some overlapping factors), and the return specific to these clusters can be used to explain the stock returns.

Wittman (2002) describes step-by-step process in clustering stock returns. The study analysed 91 US stocks across 10 industries, during the period November 1999 to November 2001. Unlike some of the studies mentioned here, Wittman used Euclidian distance to cluster stocks, instead of correlation between stocks. Although Wittman used only Hierarchical Agglomerative algorithm, he experimented with different linkage schemes (Single Link, Complete Link, Average Link and Ward's Method) that can be used in the clustering algorithm. The study also compares the quality of clusters formed using different methods by using measures such as purity, entropy and cohesion.

Wittman’s idea behind clustering is that perception moves the market and therefore, stocks should be clustered based on their behavior instead of grouping them based in standard industry classification systems. Although Wittman's belief that markets are inefficient, which is implied when he states that market is dominated by "ill-informed" participants, the clusters he obtained made more economic sense as compared to corresponding industrial classifications. To cite an interesting example, WebMD, an online prescription services that is technically a healthcare company, was classified as an internet stock. The author also makes a point that volatile stocks with little co-movement with any established group can be safely classified in a one single cluster. It is intuitively appealing to say that it would make other groups more homogeneous than they would have been with the volatile stocks.

2.b General overview of factor models

I briefly describe the construction and characteristics of factor models to show how cluster based factor model fits in the overall scheme. Factor models are used to decompose the cross-section of stock returns into pervasive factor-related components and a stock-specific component. A general form of factor model takes the form

$$R = \alpha + BF + \varepsilon \qquad (2.1)$$

where, R is n × 1 vector of returns on n stocks, F is k × 1 vector of returns on k factors, B is n × k matrix of factor loadings of n stocks on k factors, α is n × 1 vector of coefficients and ε is n × 1 vector of stock specific returns. The coefficient vector α can be defined such that expectations of stock specific return is zero. A strict factor model would assume zero correlation between the cross-section of asset specific returns. However, it is understandable that there will always be some kind of correlation between asset specific returns. Approximate factor models allow to accommodate this assumption, which ultimately leads to the understanding that asset specific risk can be diversified in a well-diversified portfolio.

The three often used categorization of factor models **[**See Connor and Korajczyk (2007)] to explain stocks returns are macro-economic factor model, statistical factor model and characteristic-based factor model.

2.b.i Characteristic-based factor model: In characteristic-based factor decomposition, factor loadings are associated with security characteristics such as book-to-market ratio, size or a security's classification by its country or industry. The three-factor model developed in Fama & French (1993) falls in this category.

The three-factor model identifies value and size as two sources of systematic risk in addition to the market risk. To obtain the associated factor returns, first the stocks are grouped in portfolios as per their book-to-market ratio and market capitalisation. The difference in return of top and bottom portfolio is considered as factor return. The factor betas are obtained by regressing time series of stock returns on the three factor returns.

Jagadeesh and Titman (1993) and Caharat (1997) identifies momentum (difference in return between outperformers and underperformers) as another source of systematic risk, and show that this factor explains variation in cross-section of stock returns in addition to the variation explained by Fama & French factors.

Factor models constructed using industry and country classification of stocks fall in the same category. This model is explained in detail later in the section 2.3.

2.b.ii Macroeconomic factor model: Factor models such as single index model in Sharpe (1970) and multifactor index model in Ross (1976) fall in this category. In this case, the factor returns are observable and time series of stock returns are regressed on factor observations using multivariate regression model to obtain factor betas.

In single factor market model, excess stock return is regressed over excess market return to obtain market beta. A broad equity market index can be considered as a proxy for the market.

Multifactor macroeconomic factor model developed by Ross uses surprises in macroeconomic variables (such as inflation and industrial production) as factors. It should be noted that these surprises are obtained as residuals after removing the interdependence of macroeconomic variables. The factor loading are obtained by regressing excess stock return on macroeconomic surprises.

Academic research [for example see Connor (1995)] confirms that macroeconomic factor models don't fit well in the observed data as compared to statistical factor models. However, macroeconomic factors models are more intuitively appealing that their statistical counterpart since factors have interpretations in terms of economic or financial theory.

2.b.iii Statistical factor model: The statistical factor models are based on the assumption that there are hidden structures in the time series of stock returns data. The model aims to identify few implicit factors, which can explain most of the variation in correlation or covariance matrix of stock returns.

The general form of statistical factor model assumes that return on n stocks at time t can be expressed as a linear function of k unobserved variables and an error term associated with each stock return. These factors can be derived using orthogonal techniques such as principal component analysis and maximum likelihood method. Each one of the orthogonal technique has its own advantages and disadvantages.

2.c. Country-Industry based factor model: I provide here a detailed discussion of factor models constructed using country and (or) industry classification for two reasons. First, the very purpose of developing a cluster-based factor model is to provide a better understanding of sources to risk and return then that provided by country-industry based factor model. Second, the structure of factor model I use is similar to the factor model to decompose international stock returns into country and industry effects.

The debate on relative importance of country or industry factors in international stock return is decades long. Lesssard (1974) showed that country effects are dominant than industry effects and therefore, diversification among countries makes more sense than diversification among industries. Lessard's results are remarkable because it says that even if one were to diversify among countries but in the same industry, it would make more sense than diversification across industries in different countries. Lassard uses 16 national and 30 international indices to reach this conclusion. Solnik (1974) also suggest that country allocation has better diversification potential than industry allocations

Heston and Rouwenhorst (1995) reach at similar conclusion but they propose a model for extracting country and industry effects, which since then has become the standard model for examining the country and industry effect (I use a similar model in my study). The regression model used by the authors has the form

$$R\,(j,t) = A\,(t) + \sum_{i=1}^{N_i} L_{i,t}\ .D_j\ + \sum_{k=1}^{N_k} L_{k,t}\ .D_j\ + \varepsilon\,(j,t) \qquad (2.2)$$

where, where R (j,t) is the return on stock j at time t, N_i and N_k are the number of industries and countries respectively, D_j is the dummy variable which is equal to 1 if stock j belongs to industry i or country k, else set as 0; for i = 1,…,N_i and k = 1,...,K_i. The unknowns in the equation $L_{i,t}$ and $L_{k,t}$ are return on industry and country factors, respectively, at time t. The other unknown A (t) is return common to all securities at time t. ε (j , t) is return specific to stock j at time t.

The regression equation effectively models the returns on stocks as

$$R\,(j,t) = \alpha_t + L_{i,t} + L_{k,t} + e\,(j,t) \qquad (2.3)$$

where, α_t is the common factor, $L_{i,t}$ and $L_{k,t}$ are industry and country effects, respectively, and e (j,t) is the stock specific effect.

It can be observed that the regression equation suffers from a dummy variable trap since it has 1/0 dummies and a constant. To correct for the problem of multi-collinearity, the authors imposed constraints that both the sum of country factors and the sum of industry factors should be zero.

This means that the regression equation constrains the least square instead of using ordinary least squares to obtain the country and industry effects.

This also changes the interpretation of country and industry effects. Now $L_{i,t}$ will be interpreted as industry effect in excess of market index of a portfolio, which has country composition same as that of the market index. Similarly, $L_{k,t}$ is now interpreted as country effect in excess of market index of a portfolio which has industrial composition same as that of the market index.

Cavaglia et.al (2000) and Hamelink et. al (2001) use the same model but reach a completely different conclusion. The authors analysed the profitability of two strategies using mean absolute deviation proposed by Rouwenhorst and concluded that industry based strategy is more profitable than country based strategy. Hamelink et. al also use simple momentum strategies to analyse the profitability of industry and country based approach and reached at same conclusion. The authors used different approaches to understand the diversification benefits of country and industry based approaches. They conclude that industry based allocation results in larger diversification benefit than country based allocation.

3. Data and Methodology

3.a. Data

I analyse the constituents of STOXX Europe 600 Index for the five year period from January 2007-December 2011. It has 600 stocks from European region which represent large, medium and small capitalisation companies. The index covers 19 industries as per Industry Classification Benchmark. Data for few stocks were not available for entire five-year period, and therefore there were removed from the analysis. In the end, the sample consists of 568 stocks.

I consider weekly returns for the purpose of analysis, which means 260 data points over a period of five years. The period of analysis could have extended but by analyzing data over 2007 to 2011, I tradeoff between capturing the latest trend and including some noise.

I could have included stocks from regions outside Europe, but I take STOXX Europe 600 to reduce computations relative to exchange rate risk in calculating returns on global equity portfolio. Also, by considering STOXX Europe 600, I try to avoid size bias by selecting a group of stocks that represent companies across large, medium and small market capitalisation. The data has been retrieved from DataStream.

3.b. Construction of factor models

A multi-factor model represents stock return in terms of sum of product of N factor returns and factor loadings plus noise. I compare two multi-factor models.

The first factor model is an industry based factor model that decomposes equity return into industry components. Formally, it can be written as follows:

$$R\,(j,t) = A\,(t) + \sum_{i=1}^{N_i} L_{i,t}\ .D_j\ +\ \varepsilon\,(j,t) \qquad (3.1)$$

where R (j,t) is the return on stock *j* at time *t* and N_i is the number of industries. D_j is the dummy variable which is equal to 1 if stock *j* belongs to industry *i,* else set as 0; for i = 1,…,N_i. The unknowns in the equation $L_{i,t}$ are return on industry factors at time *t*. The other unknown A (t) is return common to all securities at time *t*. ε (j,t) is return specific to stock *j* at time *t*.

The equation (3.1) effectively models the stocks return as

$$R\ (j,t) = \ \alpha_t + \ L_{i,t} + \ e\ (j,t) \qquad (3.2)$$

where, α_t is the common factor, $L_{i,t}$ is industry effect to which the stock belongs and e (j,t) is the stock specific effect.

I use cross-sectional regression each week over the period January 2007-December 2010, to obtain time series of returns on industry factors. To deal with perfect multicollinearity in model (1), or to ensure that A (t) has no exposure to any other industry factor I impose the constraint:

$$\sum_{i=1}^{N_i} W_{i,t}\ .L_{i,t} = 0 \qquad (3.3)$$

Here, $W_{i,t}$ is the weight of industry *i* at time *t*. Because of the unavailability of industry weights at the end of every week, I assume all industries are equally weighted. Note that the construction of model (3.1) is similar to that used in Heston and Rouwenhorst (1995).

The model (3.1) has limitations to the extent that it does not consider the influence of country, style and size factor in determining international stock returns.

I compare the performance of model (3.1) with another multifactor model constructed using clustering techniques. This model can be formally written as

$$R\ (j,t) = \ A\ (t) + \sum_{i=1}^{N_c} L_{c,t}\ .D_j \ + \ \varepsilon\ (j,t) \qquad (3.4)$$

where R (j,t) is the return on stock *j* at time *t* and N_c is the number of clusters. D_j is the dummy variable which is equal to 1 if stock *j* belongs to Cluster *c*, else set as 0; for c = 1,…,N_c. The unknowns in the equation $L_{c,t}$ are return on cluster factors at time *t*. The other unknown A (t) is return common to all securities at time *t*. ε (j,t) is return specific to stock *j* at time *t*.

The equation (3.4) effectively models the stocks return as

$$R\ (j,t) = \ \alpha_t + \ L_{c,t} + \ e\ (j,t) \qquad (3.5)$$

where, α_t is the common factor, $L_{c,t}$ is cluster effect to which the stock belongs and e (j,t) is the stock specific effect.

As described earlier, I use cross-sectional regression each week to obtain time series of returns on cluster factors. To deal with perfect multicollinearity in model (3.4), or to ensure that A (t) has no exposure to any other cluster factor I impose the constraint:

$$\sum_{i=1}^{N_i} W_{c,t}\ .L_{c,t} = 0 \qquad (3.6)$$

Here, $W_{c,t}$ is the weight of cluster *c* at time *t*. I assume all clusters are equally weighted.

3.c. Formation of clusters

It should be emphasized here that the idea of this study is not to engage in data snooping to obtain clusters that best describes the cross-section of international stocks returns. But the study aims to identify economically meaningful sources of risks in international stock returns, and tests whether clustering can do so.

Therefore, I restrict myself to common techniques used in clustering stocks. However, several points needs to be considered if clustering is applied to stock returns.

Based on Larose (2005), I identify three steps in clustering stocks. Note that clustering tries to minimize within cluster distance and maximize in between cluster distance. Therefore, I first describe consideration in selection of a measure of distance. Second, a linkage scheme should be defined. It involves defining the distance between the clusters. For example, one can ask the question "Will two clusters be considered close based on the distance between their nearest objects or the farthest objects?" Third, a clustering algorithm needs to be chosen.

3.c.i. Choice of distance measure: There is no single choice of a measure of distance but it should satisfy few basic characteristics. The distance d (i,j) between two stocks i and j should be defined in such a way that

$$d(i,j) > 0\ , for\ i \neq j \qquad (3.7)$$

$$d(i,i) =\ 0\ , for\ all\ i's \qquad (3.8)$$

$$d(i,j) = d(j,i)\ , for\ all\ i's\ and\ j's \qquad (3.9)$$

$$d(i,j) \leq d(i,k) + \ d(j,k)\ , for\ i \neq j \neq k \qquad (3.10)$$

Distance measures which satisfy criteria (i) to (iv) are called metric distances. One such metric distance takes a general form

$$d(i,j) = \left(\sum_{t=1}^{T} |i_t \ - \ j_t|^a \ ) \right)^{1/a} \qquad (3.11)$$

where, i_t and j_t are observations for the variable i and j, respectively at time t = 1,2,...,T. Some common distance measures of this form are Manhattan distance (a = 1) and Euclidean distance (a = 2).

A drawback of equation (3.11) is that its value depends on scale. To address this issue, the data can be normalized. Normalization can achieve another purpose too. Wittman (2002) suggests z-score cross-sectional normalization of data to remove the impact of market factor from the stock return.

The weekly stock return data I use to form cluster is normalized cross-sectionally. The normalized return R_j for a stock j at time t can be written as

$$R_{j,t} = \frac{R_{j,t} - \ \mu_t}{\sigma} \qquad (3.12)$$

where, μ_t and σ are mean and standard deviation of cross-sectional return of all the stocks at time t.

The distance measure in equation (3.11) suffers from another drawback. It assumes that i_t and j_t are observed at the same time, but in reality stocks prices we use in analysis might have been observed at different moments during the trading hours. Focardi (2004) suggests time wrapping to deal with this issue. However, to circumvent this additional process in the treatment of data, I didn't use the distance measure suggested in equation (3.11) in formation of cluster for factor models.

Instead, for factor model described in equation (3.4) I use correlation coefficient between the normalized time series of stock returns as a measure of distance.

It should be noted that as correlation can vary from -1 to +1, it violates the condition (3.7) and can't be used as a measure of distance. As suggested in Focardi (2004), I define a new measure of distance d(i,j) by modifying the correlation coefficient $\rho_{i,j}$ as

$$d\ (i,j) = \sqrt{2 \times \left(1 - \rho_{i,j}\right)} \qquad (3.13)$$

The function mentioned above has a range [0 2], and can now be used as a measure of distance for the purpose of this study.

3.c.ii Choice of linkage scheme: I experiment with three linkage schemes--single linkage, complete linkage and average linkage--in forming clusters using hierarchical agglomerative clustering. Single linkage defines distance between two clusters based on distance of two most similar members (one from each cluster). A drawback of this linkage scheme is that it sometimes cluster heterogeneous records. Complete linkage defines distance between two clusters based on distance of two most dissimilar members (one from each cluster). In this case few homogeneous stocks might be left out of the cluster. The Average linkage is more conservative than the two schemes discussed above and it is based on average distance of all objects in one cluster with all the objects in a different cluster.

For the purpose of using clusters in factors model, I use Average linkage scheme. Section 3.4 provides summary of various experiments with linkage schemes.

3.c.iii Choice of clustering algorithm: I use hierarchical agglomerative algorithm to cluster stocks. There are two straightforward reasons for this. First, it is one of the simplest algorithms and it provides flexibility to define distance. Second, abundant research in the past has establish hierarchical structures in the stock returns. So, instead of trying new algorithm and to avoid being accused for data snooping, I prefer using hierarchical agglomerative clustering.

Hierarchical agglomerative clustering treats each object as one single cluster and then associates two most similar clusters in each step. The cluster formation can be tracked and one can specify the number of clusters in which all the objects needs to be grouped into.

4. Empirical Results

4.a Cluster Summary

I experiment by combining hierarchical agglomerative clustering with Euclidean and Correlation (equation 3.13) distance measure. For each of the distance measure, I used all the three linkages schemes. This results in six different ways to cluster stocks.

To identify the method that results in best set of clusters, I calculate Cophenetic Correlation Coefficient for each of them. The results are reported in Exhibit 1.

Cophenetic Correlation Coefficient measures the similarity within a cluster and dissimilarity between the clusters. In general, the higher the coefficient the better the set of clusters.

Exhibit 1

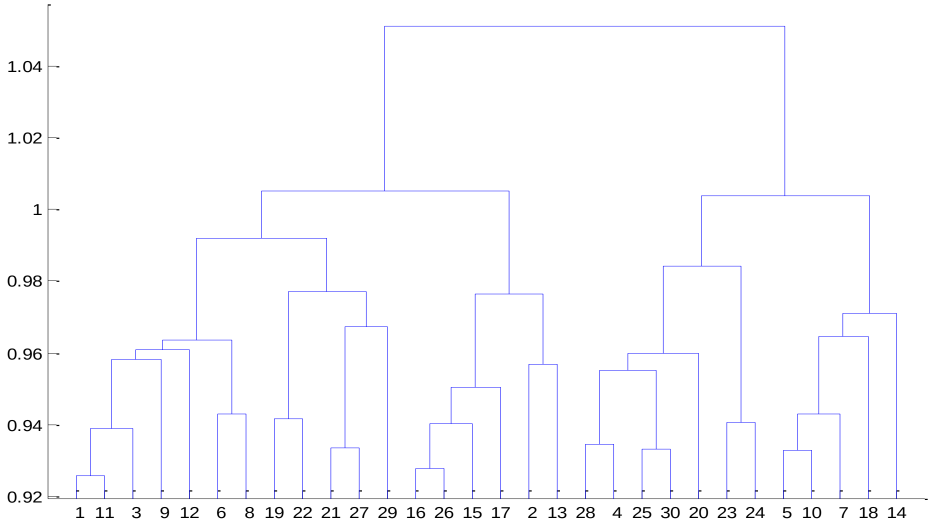
Distance measure	Linkage Scheme	Cophenetic Correlation Coefficient
Euclidean	Single	0.8864
	Complete	0.5863
	Average	0.6108
Correlation	Single	0.1014
	Complete	0.5005
	Average	0.9183

The results suggest that it is best to combine correlation distance measure with average linkage scheme as it has the highest Cophenetic Coefficient of 0.9183. The result simply validates the use of Correlation as the most appropriate measure of distance. Also, one would like to use Average linkage scheme as it is not an extreme measure like Single and Complete Linkage scheme. Therefore, the natural choice to form clusters is hierarchical agglomerative clustering, using Correlation as the distance measure and Average as the linkage scheme.

It is worth mentioning that clusters formed by combining Euclidean distance with Single linkage also results in high Cophenetic Coefficient. However, as noted above, it is inappropriate to use Euclidean measure as the stock prices used in calculating the distance may not be sometimes observed at the same time.

Another important choice to be made is the number of clusters. It is possible to track association of stock into clusters at each stage and stop clustering at a point when dissimilar stocks starts getting clustered together. However, deciding which stocks are similar involves some amount of subjective judgment, and therefore, I simply decided to form as many clusters as the number of industries. Therefore, I programmed my algorithm to produce 19 clusters. The dendrogram tree that is obtained as a result of clustering is presented below in Exhibit 2.

Exhibit 2
Dendrogram for clusters used in factor model
Distance measure: Correlation
Linkage Scheme: Average linkage
Cophenetic Correlation Coefficient: 0.9183



4.b Reclassification of stocks from industries to clusters

It is important to analyse whether the resulting clusters are backed by economic intuition. Although mapping the movement (reclassification) of all the 568 stocks from the standard industry classification to clusters can be done, I highlight only a few movements to emphasize the intuition behind the formation of clusters.

In Exhibit 3 below, it can be noticed that largest industrial classification is Industrial Goods and Services, with 90 companies. Obviously, we would not expect all the companies in this group to behave similarly and share same risks. Therefore, it is better to identify different groups within the industrial classification which share common risks. Clustering does exactly that.

I find that one third of the companies in Industrial Group and Services were regrouped into the cluster which has cyclical stocks such as auto, construction and material and basic resources. Another one-third of the companies were regrouped into cluster which has defensive stocks such as healthcare, food and beverages and utilities. The rest were classified into several other cluster, nature of which can't be verified. The result is interesting and intuitive at the same time.

We also see that Banks, Financial Services and Insurance as different industries as per ICB classification. However, upon clustering I find that most of the companies in these groups were classified into one cluster (Cluster 4). This reclassification is again logical.

Exhibit 3.

The industrial classification (ICB) of stocks as on 31 December 2011				Classification of stocks using clustering techniques as on 31 December 2011			
Industries	No. of stocks	Annualised Mean Return (%)	Annualised Standard Dev. (%)	Clusters	No. of stocks	Annualised Mean Return (%)	Annualised Standard Dev. (%)
Automobile and parts	15	5.76	20.50	Cluster 1	7	-5.87	34.48
Banks	45	16.31	32.50	Cluster 2	6	14.06	29.15
Basic Resources	26	6.27	27.04	Cluster 3	10	-7.68	37.90
Chemicals	22	-3.90	24.45	Cluster 4	50	13.01	28.87
Const. and Materials	24	-15.71	19.69	Cluster 5	9	18.84	32.27
Financial Services	28	-7.85	22.52	Cluster 6	164	6.19	27.75
Food and Beverages	27	-2.86	20.80	Cluster 7	6	6.57	26.30
Healthcare	35	-13.77	14.12	Cluster 8	5	12.95	26.50
Ind. goods and Services	90	-19.05	14.80	Cluster 9	19	12.89	30.70
Insurance	29	-2.50	20.04	Cluster 10	132	13.86	30.03
Media	27	2.70	21.90	Cluster 11	9	4.53	28.31
Oil and Gas	32	-3.30	16.53	Cluster 12	5	7.27	28.28
Personal and Hou. Goods	30	-1.94	23.48	Cluster 13	8	19.21	29.90
Real Estate	23	16.92	19.55	Cluster 14	81	14.34	43.98
Retail	24	-2.84	22.26	Cluster 15	5	1.12	27.57
Technology	26	0.09	17.43	Cluster 16	16	10.33	29.71
Telecommunications	19	5.41	19.38	Cluster 17	17	12.49	28.59
Travel and Leisure	21	-12.95	12.89	Cluster 18	15	3.94	30.51
Utilities	25	0.93	21.46	Cluster 19	4	2.03	23.91
Average Absolute Value		7.42	20.60			9.85	30.25

Cluster 6 presents another interesting example. This clusters effectively regrouped companies from relatively stable industry into one group. For example, most of its constituents are from industries such as food and beverage, healthcare, personal and household goods and utilities. One would expect relatively stable companies to behave similarly in the last five years during which the stock markets experienced high volatility.

Obviously, every cluster can't be backed by economic intuition as there must be some overlapping factors.

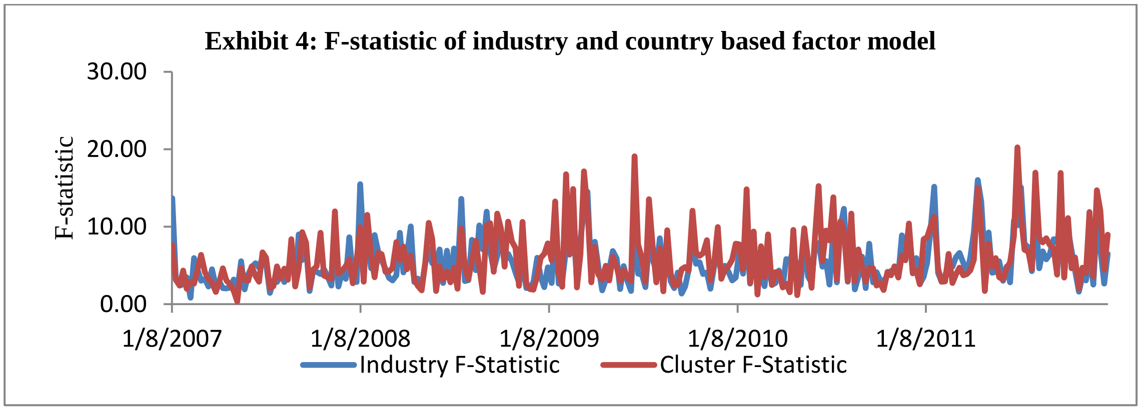
4.c. Indicators for comparing performance of factor models and results

4.c.i. Statistical significance of factor returns

I start by examining how well the two factor models fit the sample. Previous studies such as Hamelink et al. (2001) calculate cross-sectional t-statistic of individual industry and country factor returns over time, and plotted the time series of t-statistic. The authors mention that if an industry's (or a country's) t-static in all the cross-sectional regressions is continuously high and lie within a narrow range, it means the industry (country) grouping is homogenous. This allows them to compare the homogeneity of industries and countries.

Instead of comparing t-statistic, I look at F-statistic to compare industry based and cluster based factor models. F-statistic indicates how well the model fits the data, and therefore it helps in judging the overall robustness of the model. Use of F-statistic is more appropriate here than t-statistic because using the latter means comparing 19 industry t-statistic with 19 cluster t-statistic, thereby giving incomprehensible trends.

I plot the F-statistic of 260 weekly cross-sectional regressions for each industry based and cluster based factor model. A comparison between the F-statistic will tell which factor model has fitted the data well. Exhibit 4 plots the F-statistic of the both the models over time.



Although not for the entire period, but for most of the weekly periods cluster F-statistic is higher than industry F-statistic. A more clear trend emerges if moving average of F-statistic is plotted in Exhibit 5.

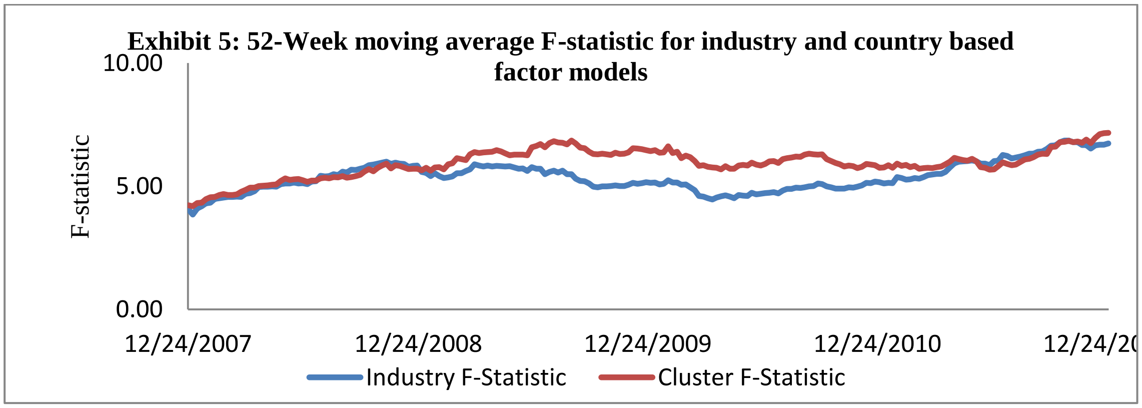


Exhibit 5 highlights that for most of the period the cluster based factor model fits the data better than the industry based factor model. The difference is more prominent during 2008-10. The prominence of cluster effect is also clear from the summary statistics provided in Exhibit 6. The average F-statistic of 260 cluster based regression models is 5.89, higher than that of industry based regression models. Also significant is the difference in p-values. For cluster based regression, the probability at which all the coefficient are zero at the same time is negligible.

Exhibit 6

	Average F-statistic	**Average p-value**
Industry based models	4.85	0.0185
Cluster based models	5.89	0.0076

4.c.ii Profitability of strategies

Determining the profitability of industry and cluster based strategies is important to compare their relative importance. Rouwenhorst (1999) proposed Mean Absolute Deviation (MAD) to understand relative importance of country and industry factors. I use the same measure to compare the profitability of industry and cluster based strategies.

MAD, as defined here, is weighted average of absolute mean industry (cluster) factor effects at a time *t*. The factor returns can be either equally weighted or market capitalisation weighted. I assume industries (clusters) are equally weighted. Formally, industry (4.1) and cluster (4.2) MAD at a time *t* can be written as:

$$MAD_{industry,t} = \sum_{i=1}^{N_i} W_{i,t-1} \, . \, |L_{i,t}| \tag{4.1}$$

$$MAD_{cluster,t} = \sum_{c=1}^{N_c} W_{c,t-1} \, . \, |L_{c,t}| \tag{4.2}$$

where, $W_{i,t-1}=1/N_i$ and $W_{c,t-1}=1/N_c$ and $L_{i,t}$ and $L_{c,t}$ are industry and cluster factor returns, respectively.

Industry MAD can be interpreted as gains from a strategy based exclusively on pure industry factor returns. The mathematical construct of MAD implies that portfolio manager has perfect insight, which means the manager takes long position in industry that will give positive return in the subsequent period and short position in industry that will give negative return in the subsequent period. For example, if there are just two industries (assume they have equal market capitalisation) and factor returns associated with them at time t are $L_{1,t}$ and $L_{2,t}$, then the return from industry based strategy or $MAD_{industry,t}$ will be $(L_{1,t} + L_{2,t})$. Cluster MAD is interpreted in the same way.

Exhibit 7 plots time series industry and cluster MAD to analyse gains from following the two strategies separately. It is noticeable that although weekly MAD is volatile, cluster MAD is higher than the industry MAD for most of the weeks. Exhibit 8 plots industry and country MAD after smoothing over the trailing 52-weeks.

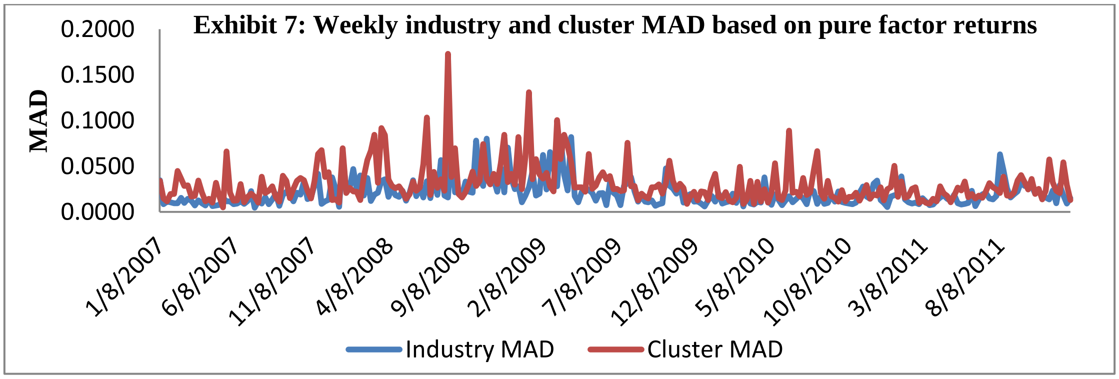


Exhibit 7: Weekly industry and cluster MAD based on pure factor returns

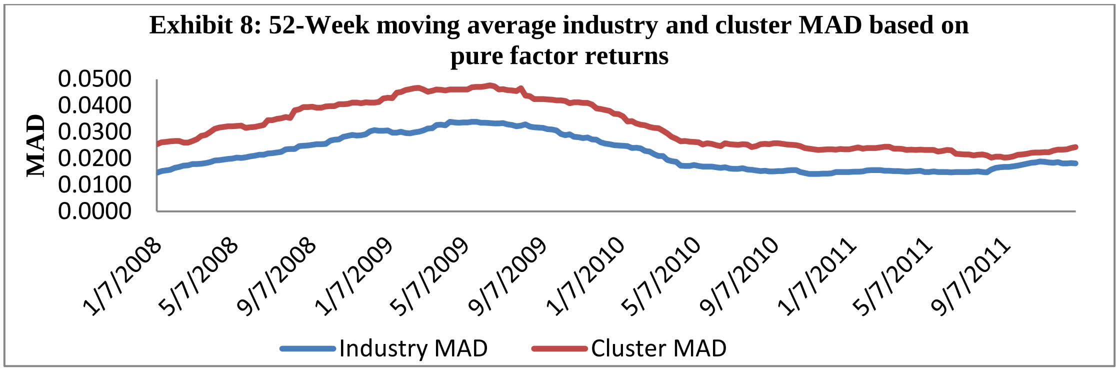
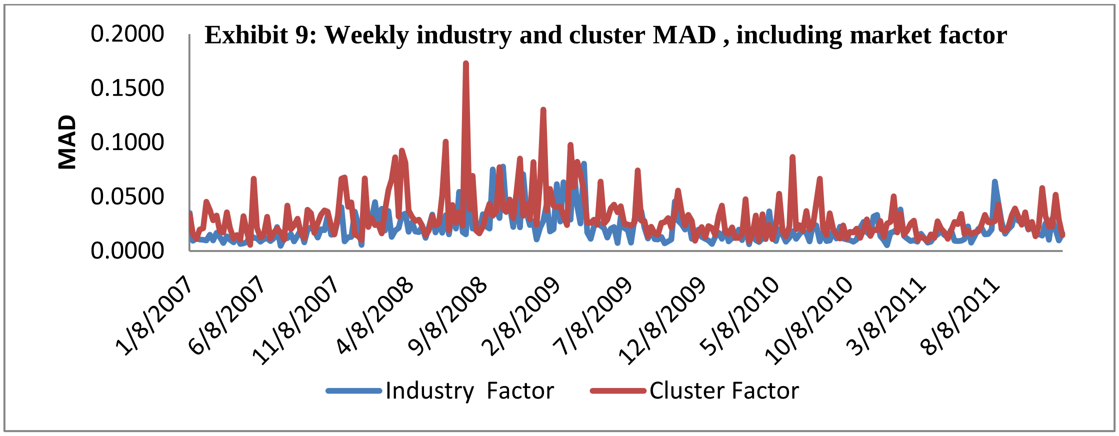


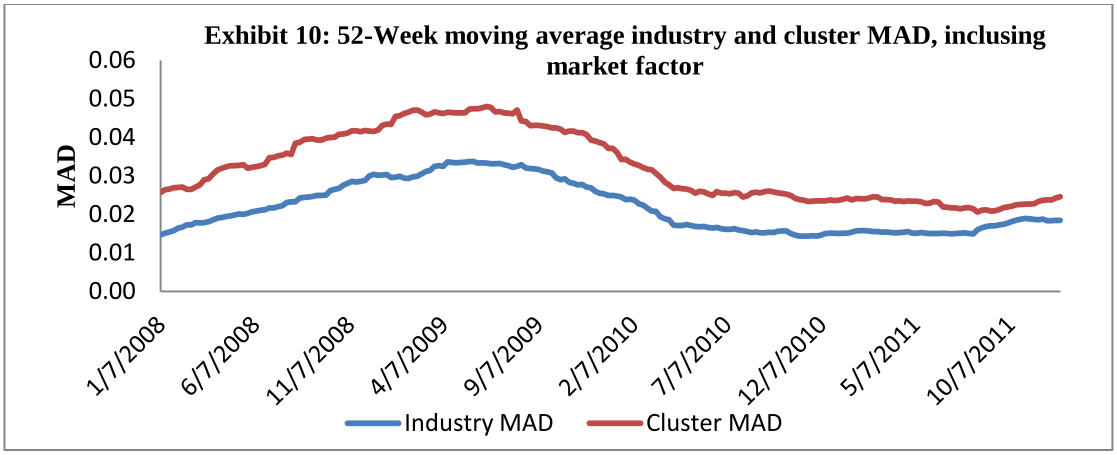
Exhibit 8: 52-Week moving average industry and cluster MAD based on pure factor returns

Exhibit 8 shows that although the difference between the magnitude of cluster and industry effect changes over time, the magnitude of cluster effect is consistently higher than industry effect. This implies that if a portfolio manager has perfect insight, he or she can gain more by allocating resources over clusters than over industries.

For exhibits 7 and exhibit 8, market factor common to all industries or countries is not included to calculate MAD. It is obvious that market factor obtained in cluster and industry based factor model may be significantly different. This could change the conclusions drawn above.

To check the robustness of the model, I include market factor return to calculate industry and country MAD. It report both weekly MAD and 52-week moving average MAD in Exhibit 9 and Exhibit 10, respectively. The conclusion drawn above remains unchanged.





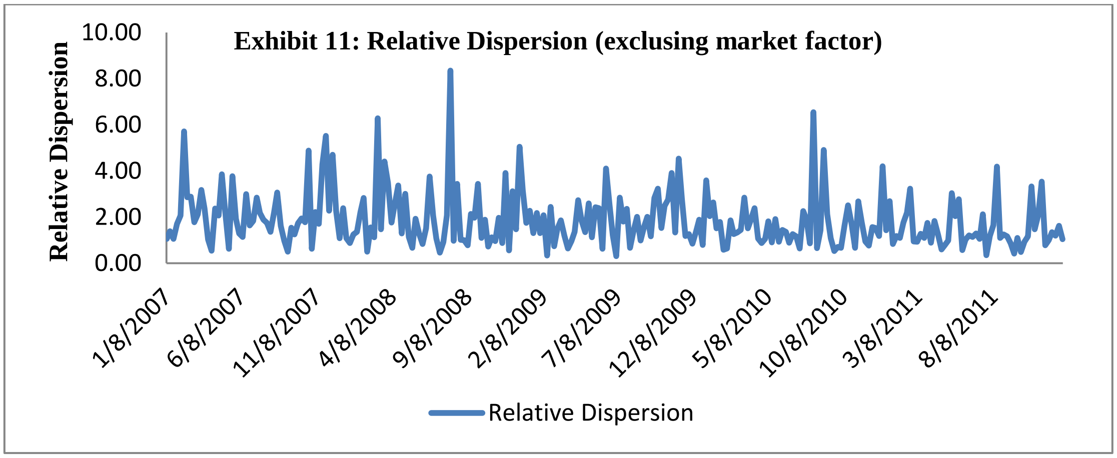
It should be noted here that in practice it is impossible to have perfect insight, and therefore replicating the performance indicated above is not possible. However, it does show that cluster based strategies have higher potential to generate returns than the industry based strategy. Techniques such as momentum strategies (Hamelink et al., 2001) and statistical measures such as Mean Square Prediction Error (West, 2006) can also be used to analyse the predictability of factor models, but these methods have been left out from the scope of this paper.

4.c.iii Diversification benefits

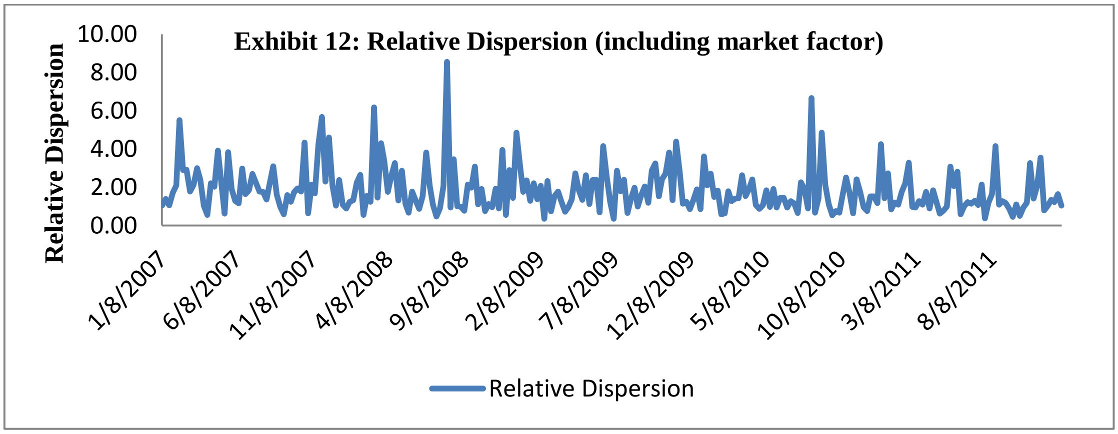
One way to compare relative benefits of industry based allocation versus cluster based allocation strategies is to analyse their diversification benefits. Diversification arises from low correlation among asset classes. Similarly, diversification benefits from industry and cluster based allocation strategies can be inferred from correlation among industry factor and cluster factor returns.

Solnik and Roulet (2000**)** suggested cross sectional dispersion (standard deviation) as an alternative to correlation. I apply this measure to analyse potential diversification benefits from industry and cluster based strategies in volatile periods. Higher cross sectional dispersion among industry (cluster) returns indicate higher diversification potential from adopting industry (cluster) based strategy.

Exhibit 11 reports relative dispersion of pure cluster and industry factor returns. Relative dispersion as defined here is cross sectional dispersion of cluster divided by cross sectional dispersion of industry. Relative dispersion higher than 1 indicates cross sectional dispersion of clusters is higher than that of industry.



The graph shows that for most of the weeks the cluster dispersion is higher than the industry dispersion. This suggests that potential diversification benefits arising from investing across clusters is higher than that from diversifying across industry. Exhibit 12 shows relative dispersion calculated after including common factor return. The conclusion is unchanged.



I also analyze diversification benefits of industry and cluster based strategy in the most volatile periods. I list down 52 worst performing weeks for the European equity markets during 2007-2010, by taking STOXX Europe 600 as the benchmark index. Next, I calculate cross-sectional dispersion of industry factor returns and cluster factor returns during these weeks.

I find the number of months in which relative dispersion is higher than one. Exhibit 13 below shows that in 41 out of 52 weeks, the relative dispersion was higher than 1. It can thus be inferred that cluster based strategy provide better diversification benefits than industry based strategy even in the most depressed periods for the equity markets.

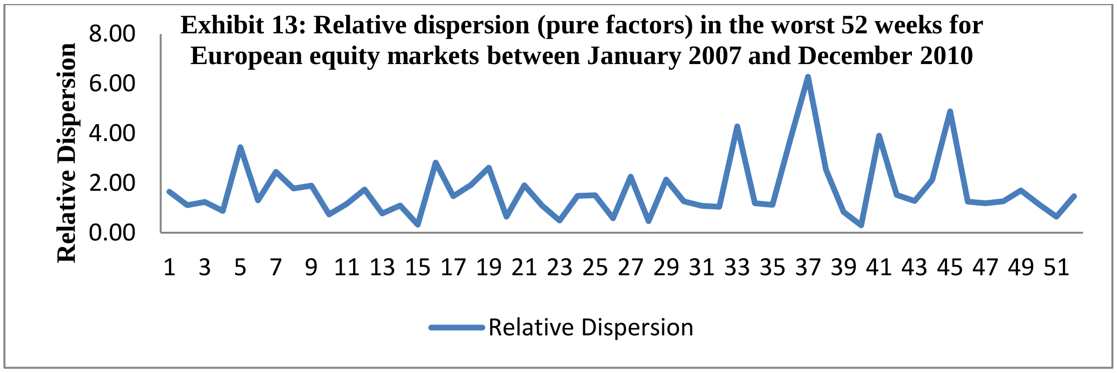


Exhibit 13: Relative dispersion (pure factors) in the worst 52 weeks for European equity markets between January 2007 and December 2010

Exhibit 14 plots relative dispersion by including common factor return in the calculation of dispersion. It again suggests that when investors need benefits of diversification the most, cluster based strategy provide better diversification potential than industry based strategy.

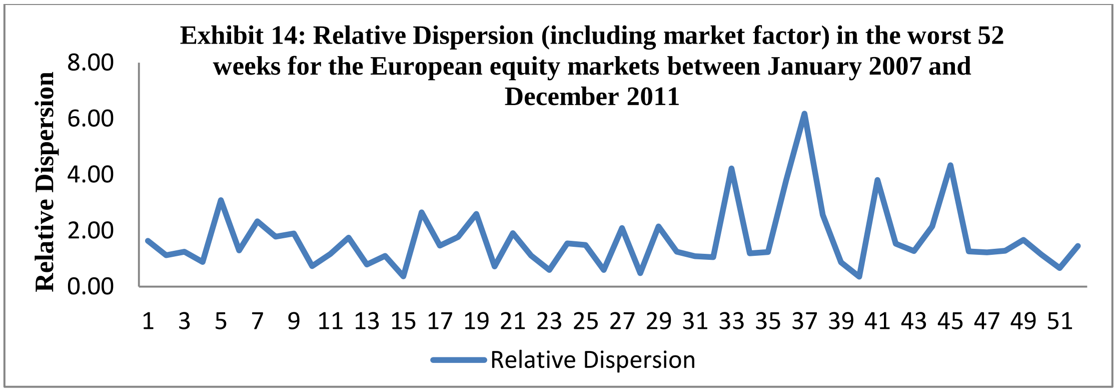


Exhibit 14: Relative Dispersion (including market factor) in the worst 52 weeks for the European equity markets between January 2007 and December 2011

5. Conclusion

I construct a factor model assuming that stocks returns can be expressed as cluster based effects. I compare it with a factor model based on industry classifications. The empirical evidence suggest that cluster based factor returns are more homogeneous and significant than industry based factor returns. Also, with perfect insight, cluster based strategies are more profitable than industry based strategies. Further, it find that diversification benefits arising from cluster based strategies are higher than industry based strategies. Finally, cluster based strategy provides better diversification benefit than industry based strategy in volatile times, when diversification is needed the most.

These results have implications for global equity portfolio managers who use industry based factor models. Empirical evidences presented here suggest that it is profitable to build factor model by clustering stocks that behave similarly, rather than building a factor model using the standard industry classification.

6. REFERENCES:

Arnott, Robert D., 1980, "Cluster Analysis and Stock Price Comovement," Financial Analyst Journal, 56-62

Beckers, Stan, Gregory Connor, and Ross Curds, 1996, “National versus Global Influences on Equity Returns,” Financial Analysts Journal, 31-39

Carhart, M. M., 1997, "On Persistence in Mutual Fund Performance," Journal of Finance, 57-82

Cavaglia, Stefano, Christopher Brightman, and Michael Aked, 2000, "The Increasing Importance of Industry Factors," Financial Analysts Journal, 41-53

Connor, G., 1995, "The Three Types of Factor Models: A Comparison of their Explanatory Power," Financial Analysts Journal, 42-46

Connor, Gregory and Korajczyk, Robert A., 2009, "Factor Models of Asset Returns," Encyclopedia of Quantitative Finance, Wiley

Fama, E. and K.R. French,1993, "Common Risk Factors in the Returns on Stocks and Bonds," Journal of Financial Economics, 3-56

Farrell, Jr., James L., 1974, "Analyzing Covariation of Returns to Determine Homogeneous Stock Groupings," Journal of Business, 186-207

Focardi, Sergio M., 2001-04, "Clustering Economic and Financial Time Series: Exploring the Existence of Stable Correlation Conditions," Discussion Paper, The Intertek Group

Grinold, Richard, Andrew Rudd, and Dan Stefek. 1989, “Global Factors: Fact or Fiction?,” Journal of Portfolio Management, 79-89

Hamelink, Foort, Helene Harasty, and Pierre Hillion, 2001, "Country, Sector or Style: What Matters Most When Constructing Global Equity Portfolios? An empirical investigation From 1990-2001," FAME Working Paper No. 35.

Heston, Steven L., and K. Geert Rouwenhorst, 1995, "Industry and Country Effects in International Stock Returns," Journal of Portfolio Management, 53-58

Jagadeesh, N. and S. Titman,1993, "Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency," Journal of Finance, 65-91

Larose, Daniel T., 2005, "Discovering Knowledge in Data: An Introduction to Data Mining," John Wiley & Sons, Inc.,147-162

Lessard, Donald, 1974, “World, National, and Industry Factors in Equity Returns,” Journal of Finance, 379-391

Mantegna, R.N., 1999, "Hierarchical Structure in Financial Markets," European Physical Journal B, 193-197

Ross, S., 1976, “The arbitrage theory of capital asset pricing,” Journal of Economic Theory 13, 341-360

Rouwenhorst, K. Geert, 1999, "European equity markets and the EMU," Financial Analysts Journal, 27-34

Sharpe, W.F., 1970, "Portfolio Theory and Capital Markets," McGraw-Hill, New York.

Solnik, Bruno, 1974, “Why not Diversify Internationally Rather than Domestically,” Financial Analyst Journal, 48-54

Solnik, Bruno, Jacques Roulet, 2000, "Dispersion as Cross-Sectional Correlation", Financial Analysts Journal, 54-61

West , Kenneth D., 2005, "Forecast Evaluation," Handbook of Economic Forecasting, Elsevier B.V

Wittman, Todd, 2002, "Time-Series Clustering and Association Analysis of Financial Data," CS 8980 Project