

CATEGORIZING MUTUAL FUNDS USING
CLUSTERS

by

Achla Marathe
Los Alamos National Laboratory,
Mail Stop B265,
Los Alamos, NM 87545.
email: achla@lanl.gov
Phone: (505)-667-9034

and

Hany A. Shawky
University at Albany,
School of Business,
Albany, NY 12222.
email: h.shawky@albany.edu
Phone: (518)-442-4921

ABSTRACT

We classify mutual funds using cluster analysis and examine whether the categories created by clusters are the same as those assigned by the investment objectives. Since mutual funds are usually classified based on their investment objectives, clustering funds based on their financial characteristics, rather than their investment objectives could explain why some mutual funds do not perform according to their investment objectives. After clustering we find that some of the investment categories are insignificantly different from others in terms of their financial characteristics. Maintaining more investment categories than necessary causes inefficiency for the financial analysts and mutual fund managers because the benefits of diversification can not be achieved by investing in different categories of funds.

Categorizing Mutual Funds using Clusters

1 Introduction

It is important for a mutual fund investor to know whether the funds' performance is consistent with its stated objective. An investor needs to be able to select funds that best suit his financial needs, risk levels and help him anticipate the future flow of returns from his investment. Also, for an investor to manage risk efficiently, through optimal diversification, it is imperative that the mutual funds are classified in their proper investment category. Misclassification of mutual funds can lead investors to allocate their resources into funds whose risk and return characteristics do not match their expectations. Lack of knowledge about the future risk-return pattern can result in suboptimal investment decisions and hence suboptimal consumption/investment choices.

The mutual fund managers have several incentives to misclassify their fund's investment objective. Foremost is to get higher ranking among its peers. Given that most of the mutual fund managers get compensated based

on their fund's performance and ranking, they have an incentive to intentionally misclassify their fund's objective. If a fund exposes itself to more risk than its objective claims, it is likely to have higher returns than its peers in the same category and hence a better ranking than the rest. Brown, Harlow and Starks (1996) show that managers who performed relatively poorly in the first half of the year tend to take on more risk in the second half of the year to respond to their incentive structures.

Several magazines and publications list the top performers in different categories and give free publicity to the top performers. This provides an added incentive for the smaller companies with tiny advertising budgets to misclassify their funds to take advantage of this free publicity. In order to have a fair comparison of performance and ranking, and to avoid comparing apples with oranges, we need to further explore and perhaps more objectively identify the right classification of individual mutual funds.

Several studies have examined the relationship between funds' stated objective and their measures of risk and return. For instance, diBartolomeo and Witkowski (1997) use a return based methodology developed by Sharpe (1992) to classify mutual funds. Their results show that 40% of all equity funds are misclassified. Based on Monte Carlo simulations, they find that misclassification has a significant impact on the investor's ability to diversify his portfolio of mutual funds. McDonald (1974) examined the overall

performance of a sample of mutual funds relative to their self-declared objectives and found a positive relationship between investment objectives and measures of risk.¹ Martin, Keown and Farrell (1982) examined a sample of mutual funds representing five investment objectives and found definite differences in the variability of the funds in alternative classifications.² We classify mutual funds using cluster³ analysis and examine whether the categories created by clusters are the same as those assigned by the self-declared investment objectives. Since mutual funds are usually classified based on their investment objectives, clustering funds based on their financial characteristics, rather than their investment objectives could explain why some mutual funds do not perform according to their investment objectives. After clustering, we find that 43% of the mutual funds are misclassified.⁴ In many instances self-declared categories of funds are indistinguishable from one another when objective financial characteristics are used to classify them.

This paper is organized as follows. Section 2 explains the data used for clustering. Section 3 describes the k-means algorithm and shows the clustering results when all the 28 variables are used. Section 4 confirms the robustness of clusters and the final section concludes the paper.

2 Data

The data used in this study is obtained from *Morningstar* for the year 1995. It uses 904 different funds having seven different investment objectives. The categories of investment objectives with adequate sample size include World Wide Bonds, Growth, Small company, Municipal NY, Municipal CA, Municipal State and Municipal National. Based on the availability and completeness of the data we selected twenty eight financial variables for each of the funds.⁵ The variables used to perform the cluster analysis are listed in the Appendix.

In order to effectively cluster the data, all the variables are normalized so that each resulting column has mean zero and variance one. The k-means clustering technique applied in this analysis uses Euclidean distance. Euclidean distance between two objects i and j can be measured as

$$d_{ij} = \left[\sum_{k=1}^p (X_{ik} - X_{jk})^2 \right]^{1/2}$$

which is not scale invariant. Hence, when computing distances between objects, the raw data should be appropriately scaled to preserve distance rankings.

3 Cluster Analysis of the Data

3.1 The k-means Clustering Algorithm

Clustering involves dividing the set of data points into non-overlapping homogeneous groups or clusters of points, which are internally cohesive. If the objects can be represented by points in Euclidean space, the k-means criterion can be used. k-means is an iterative relocation algorithm, where an initial classification is modified by moving objects from one group to another such that it minimizes the within group sum of squares.

The k-means algorithm is set up in the following way.⁶ Initial reference points, which may or may not be the centroid or mean are chosen and all the data points are assigned to clusters. k-means then uses the cluster centroids as reference points in subsequent partitionings but the centroids are adjusted both during and after each partitioning. For data point x in cluster i , if the centroid z_i is the nearest reference point, no adjustments are made and the algorithm proceeds to the next data point. However, if the centroid z_j of the cluster j , the reference point is closer to data point x , then x is reassigned to cluster j , the centroids of the ‘losing’ cluster i and the ‘gaining’ cluster j are recomputed and the reference points z_i and z_j are moved to their new centroids. After each step, every one of the k reference points is a centroid or mean and that is why it is called “k-means.”

This method requires one to specify the number of clusters in advance.

Hartigan (1975) suggested the following rule of thumb to find the optimal number of clusters. If k is the result of k-means with k clusters and $k + 1$ is the result with $k + 1$ clusters, then it is justifiable to add the extra cluster when

$$\left(\frac{\sum_{i=1}^k ESS}{\sum_{i=1}^{k+1} ESS} - 1 \right) * (n - k - 1) > 10$$

where ESS represents the within sum of squares and n is the size of the data set.

3.2 Clusters in Multi-Dimensional Space

We used Hartigan (1975) rule of thumb to determine the optimal number of clusters. The result was a set of 23 clusters made in the 28 dimensional space using all the original normalized variables. The clusters divided the data set into three very distinctive groups. Group one containing the World Wide Bonds, group two, containing all the Growth and Small Company funds, and the third group having all the municipal funds.⁷ The k-means algorithm minimizes the within group sum of squares to get the best classification of the data. Table 1 shows the 23 clusters where each column shows the number of funds belonging to that category and rows show the cluster numbers. For example, row 12 can be read as cluster number 12 having one growth fund, one small company fund and two municipal national funds. Given that clusters 1, 2, 4 and 17 have only world wide funds, they can be labeled as

World Wide Bond clusters. Clusters 3, 5, 6, 7, 8, 9, 10, 11, 12, 15 and 21 are grouped as the Growth and Small Company fund clusters and cluster 13, 14, 16, 18, 19, 20, 22 and 23 as the Municipal fund clusters.

Insert Table 1 here.

The classification of the mutual funds based on the financial characteristics gives a different grouping than are given by their stated investment objectives. The new groupings can be justified by looking at the differences in mean risk and return variables of funds. Table 2 gives the mean risk and return values of the mutual funds for each original investment category and also the newly formed clusters. The risk and return variables used are the one year total return, 3 year annualized return, 5 year annualized return, 3 year standard deviation, 5 year standard deviation, alpha and beta.⁸

Lets first analyze the funds whose stated objectives are “Growth” and “Small Company”. The small company funds typically have higher risk and return than the growth funds. The 3 year and 5 year risk-return values comply with that expectation but the one year total return and beta tell a different story. The one year mean return for the growth funds is higher than the one year mean return for the small company funds. Beta, which measures the market price of risk, is lower for the small company funds than for the growth funds. Given that the two funds with different stated objectives give such mixed signals to the investor about the risk-return pattern, it may be

more appropriate to put both of them in one category.

All the municipal funds are so similar in risk-return characteristics that they almost look indistinguishable from one another. Municipal CA and State have almost identical 5 year return and very similar 3 year return. Municipal NY has higher 5 year return but lower 5 year risk than municipal CA. The spread on 3 and 5 year returns for all municipals is only 1.51 percent whereas the spread on 3 and 5 year risk is 1.47 percent. The one year returns vary in the range of 16.88 and 18.33. The alphas and betas also vary in such narrow ranges that it appears more rational to have all of the municipal funds in one category rather than four different categories. For investors who wish to diversify their portfolios of mutual funds, simpler and fewer categories of investment objectives are easy to manage. Given that the risk-return characteristics are not substantially different among the four classes of municipals, it is inefficient for the investor to analyze four different categories of mutual funds.

The three cluster categories shown in table 2 have very distinct risk-return characteristics. The growth and small company clusters have the highest one year mean return of 31.09%. The municipal and world-wide clusters have a mean return of 17.43% and 13.89% respectively. The alpha and beta are also very distinct between different categories but similar within each category. The world wide clusters have higher risk but lower returns

than the municipal clusters. This is possible given that world wide funds reflect global characteristics. The risk-return pattern across countries can be very different from the risk-return pattern within a country. Based on the above clusters we find that 43% of the funds analyzed in this study are misclassified.

Insert Table 2 here.

4 Robustness of the Clusters

We have used several different variables to measure risk and return, some of which are short term while others are long term. It is natural to expect strong correlations between the short term and long term variables. We construct pair-wise graphs of the five different return variables which are Morningstar's 3 year return, 5 year return, 1 year total return, 3 year annualized return and the 5 year annualized return. Figure 1 shows that there are high correlations between all the return variables. Removing the correlating components of the data will have the advantage of removing the redundancy in the data. This can be done by using the Principal Component technique described later in this section.

Figure 2 presents similar results pertaining to the risk variables. The pair-wise graphs represent some of the risk variables which are Morningstar's 3 year risk, 5 year risk, beta, 3 year standard deviation and 5 year standard

deviation.⁹ The strong correlation between these estimated risk variables is apparent. Figures 1 and 2 also show the correlations between the short term and long term risk and return variables.

Insert Figure 1 and Figure 2 here.

To remove redundancy in the data,¹⁰ we use Principal Components analysis. An essential feature of the Principal Component Analysis (PCA)¹¹ is that it reduces the dimensionality of a data set which consists of a large number of interrelated variables, while retaining maximum possible variation in the data set. This is done by transforming to a new set of variables, the principal components (PCs), which are uncorrelated, and which are ordered so that the first few retain the most of the variation present in all of the original variables.

The PCA is applied on the set of all 28 variables which include all of the short-term and long-term risk and return variables. The results in Table 3 show that the first 16 components explain more than 98% of the variance in the original variables. They are the linear combinations of the original 28 variables with correlating variables receiving less weights. This implies that we can leave out the rest of the 12 components without any significant loss of information. Now we use the reduced data set of the first 16 principal components to cluster the mutual funds. The robustness of the clusters can be shown if this data set also forms similar clusters.

Insert Table 3 here.

The k-means clustering technique is applied to construct 23 new clusters using the first 16 principal components. In this 16 dimensional space the k-means within group sum of squares goes down by 4%.¹² The classification of the funds given by the principal components method is shown in Table 4. This Table shows the 23 clusters where clusters 1, 2, 4 and 17 can be labeled as World Wide Bond clusters, clusters 3, 5, 6, 7, 8, 9, 10, 11, 12, 15 and 21 as the Growth and Small Company fund clusters and cluster 13, 14, 16, 18, 19, 20, 22 and 23 as the Municipal fund clusters. Note that there is no significant change in the funds categories when fewer dimensions using the principal components are used.¹³ This implies that the classification given by clustering is robust irrespective of the dimensionality of the data set used.

Insert Table 4 here.

5 Summary and Conclusions

This study explores the use of clustering technique in grouping mutual funds with different investment objectives. Cluster analysis has the inherent ability to accommodate non linearities and complex interactions among explanatory and explained variables without imposing any structural relationships. We find 43% of the mutual funds do not belong to their stated categories. The classification given by the k-means clustering is much simpler than the one

given by the funds and reported in Morningstar.

Using statistical techniques instead of using “stated objectives” to categorize mutual funds has the potential of explaining differences in the risk-return performance of the various fund categories. In addition, our analysis indicates that despite the very large number of proclaimed fund categories, they seem to behave in very similar fashion when it comes to risk and return indicators. Using clusters, the present mutual funds sample could be effectively divided into three groups. The world-wide bond group contains all the world wide funds. The growth group consists of all the growth and small company funds. This is not surprising given that most small companies are characterized by a significant growth component. Finally, the municipal funds group contains all the different kinds of municipal funds i.e. Municipal National, Municipal State, Municipal CA, Municipal NY.

The municipal funds could have been better classified if we had some state specific variables to discriminate between different state municipal funds. However, the risk-return characteristics are still too alike to justify having different categories. The growth and small company funds had good discriminating variables like risk, return, alpha, beta, percentage of stocks, bonds, cash etc. and hence, having them fall into one investment category presents strong evidence that their investment characteristics are not sufficiently different for each to belong to a different group of funds.

Thus, clustering can help simplify the classification of funds by consolidating the data based on funds' characteristics rather than stated investment objectives.

APPENDIX

The Morningstar's Variables Used to Perform the Cluster Analysis

1. Morningstar Risk (3 year)
2. Morningstar Risk (5 year)
3. Morningstar Return (3 year)
4. Morningstar Return (5 year)
5. 1 Year Total Return
6. 3 Year Annualized Return
7. 5 Year Annualized Return
8. 1994 Annual Return
9. 1993 Annual Return
10. 1992 Annual Return
11. 1991 Annual Return
12. Alpha (3 year)
13. Beta (3 year)
14. Standard Deviation (3 year)
15. Standard Deviation (5 year)
16. Income Ratio
17. Turnover
18. Potential Gain Exposure
19. % Cash
20. % Stocks
21. % bonds

22. % Preferred

23. % Other

24. Maximum Sales Charge

25. % Front Load

26. % Deferred

27. Expense Ratio

28. Net Assets \$MM

ACKNOWLEDGEMENTS

We greatly acknowledge the helpful comments and suggestions by the editor Cheng F. Lee and two anonymous referees.

NOTES

1. Positive relationship means as risk measures increased the investment objectives became more aggressive.
2. Shawky (1982) examined a comprehensive sample of mutual funds and found that although most funds had below average performance over the years, they improved the diversification of their portfolios and their risk was consistent with fund objectives. Ang and Chua (1982) examined the consistency of performance of funds with different investment objectives and found that various funds met their stated objectives but did not do it consistently.
3. According to Hartigan (1975), clustering can help in summarizing, predicting and explaining information on data based on the characteristics of clusters. Unlike regression approach, the cluster analysis does not impose any linearity restrictions or theoretical structure between the endogenous and exogenous variables.
4. The misclassification rate is based on the clusters. All the worldwide bond funds which do not fall under WW cluster are called misclassified. Given that all the small company funds and growth funds belong to the same clusters, we call this joint category “Growth”. In this case all the small company funds are declared

misclassified. Similarly, all the municipals effectively belong to only one set of clusters, so we declare muni NY, muni CA and muni State to be misclassified.

5. Variables with sparse data are not included here.
6. For more detailed analysis of the K-means algorithm, see Faber (1994) and Hartigan (1975).
7. The within group sum of squares was found to be 6964.44.
8. Morningstar computes the 3 year and the 5 year standard deviations using monthly return data.
9. While the standard deviation measure is directly calculated from the funds' monthly returns, the 3 and 5 year risk variables are calibrated measures relating to industry return variability.
10. In regression terminology, this is tantamount to the process of reducing multicollinearity between the independent variables.
11. For more detailed information on the use of principal components analysis, see Jolliffe (1986).
12. The within group sum of squares is 6715.32.

13. It should be noted that we have tried to analyze two different kinds of correlations in the data set. The first is the correlation between variables, e.g. 3 year risk, 5 year risk and 3 year return, 5 year return etc. which is analyzed using principal components. The correlating variables are converted into orthogonal variables without sustaining a significant information loss. The reduced number of principal components give equally good results in discriminating between clusters. The second is the correlation between funds which is analyzed using the cluster analysis. The funds with similar information in their financial characteristics are grouped into similar clusters.

References

- [1] James S. Ang and Jess H. Chua (1982) “Mutual Funds: Different Strokes for Different Folks?,” *The Journal of Portfolio Management*, No. 2, Winter, pages 43-47.
- [2] Keith C. Brown, W.V. Harlow and Laura T. Starks (1996) “Of Tournaments and Temptations: An Analysis of Managerial Incentives in Mutual Fund Industry,” *Journal of Finance*, vol. 51, no. 1, pages 85-109.
- [3] diBartolomeo Dan and Eric Witkowski (1997) “Mutual Fund Misclassification: Evidence based on Style Analysis,” *Financial Analysts Journal*, Sep/Oct, pages 32-43.
- [4] Faber, Vance (1994), “Clustering and the Continuous k-means Algorithm,” *Los Alamos Science*, Los Alamos National Laboratory.
- [5] Hartigan, J. A. (1975), *Clustering Algorithms*, New York: Wiley.
- [6] Jolliffe, I.T. (1986), *Principal Component Analysis*, Springer-Verlag: New York.
- [7] John D. Martin, Arthur J. Keown Jr. and James L. Farrell (1982) “Do Fund Objectives Affect Diversification Policies?” , *The Journal of Portfolio Management*, No. 2, Winter, pages 19-28.

- [8] John G. McDonald (1974) "Objectives and Performance of Mutual Funds," *Journal of Financial and Quantitative Analysis*, No. 3, June, pages 311-333.

- [9] William F. Sharpe (1992) "Asset allocation: Management style and performance measurement," *The Journal of Portfolio Management*, winter, pages 7-19.

- [10] Hany A. Shawky (1982) "An Update of Mutual Funds: Better Grades," *The Journal of Portfolio Management*, No. 2, Winter, 29-34.