Positive Portfolio Factors

Stephen J. Brown
Stern School of Business, NYU

William N. Goetzmann
Yale School of Management

Mark Grinblatt
Anderson School of Management, UCLA

First Draft: December 22, 1996

January 9, 1997

Abstract: we use an iterative relocation algorithm to identify factors in common stock returns. The benefit of the approach is that factors are portfolios of assets with non-negative weights. As a results, they are readily interpreted in terms of their characteristics of the underlying securities. The positive portfolio factors have comparatively high explanatory power in sample and out-of-sample. We find evidence of a size factor and factors identified with certain industries. Factors extracted from the mutual fund universe perform marginally better than factors from the universe of equities.

I. Introduction

It is commonly agreed that stock returns are well described by a multi-factor model, however there is little agreement regarding the exact identification of the factors. Principal components methods used by Conner and Korajczyk(1988) and Lehman and Modest(1988) and factor analysis procedures used by Roll and Ross (1980) provide factors with the greatest explanatory power, however the resulting rotations contain little intuition with respect to the factors that are "priced" in the economy. Chen, Roll and Ross (1986) show that exposures to widely discussed macro-economic variables such as innovations in inflation, the term structure and production generate positive risk premia. Recently, the most successful approach to identifying multi-factor models are methods using pre-specified rotations of stocks. Elton and Gruber (1994) create factors are the sequence of residuals obtained by successively regressing a small stock index and a bond index on the S&P 500. Fama and French (1992) and Grinblatt and Titman (1985) choose portfolios according to "fundamental" variables such as size, dividend yield and earnings-price ratio. Fama and French (1992) sort stocks in these dimensions and take differences between highest and lowest deciles to create factors. Grinblatt and Titman (1985) estimate portfolios that are maximally correlated to exposures to fundamental variables. Both find that measures and ratios commonly used to characterize the value of a security in fact also help explain out-of-sample differences in expected returns.

In this paper, we use a different approach to identify factors in common stock returns. Rather than pre-specifying factors or rotations, we allow factors to be endogenously determined. The primary benefit of our procedure is that it yields factors which have non-negative portfolio weights.

Thus factors capturing differences in risk can be mimicked by passive, positive investment portfolios. In addition we can identify the characteristics of these portfolios such as dividend yield, earnings-price ratio and size and industrial composition.

The approach we use is a variation on clustering methods employed by Elton and Gruber (1969) and applied more recently in Brown and Goetzmann (1997) to classify mutual fund managers according to style. Elton and Gruber (1969) use an iterative re-location algorithm to break the universe of stocks down into meaningful sub-sectors in order to reduce estimation error in mean-variance optimization. Although their work proceeded empirical attempts to estimate APT models by more than a decade, clustering methods have not been applied to the problem of factor identification. The reasons for this are apparent from the derivation of the APT. In a framework where factors are known and estimable, covariance of any security with the set of risk factors is sufficient to determine its expected return. As a consequence, most empirical APT analysis has focused on the covariance matrix of security or portfolio returns. Clustering takes a different tack. We begin with the assumption that approximate factors may be obtained as equal-weighted portfolios of securities. To estimate these portfolios, securities are grouped together to minimize the sum of within-group squared errors over the time period of estimation. For example, two securities whose returns nearly match each other each observation period over a given time interval will cluster together. No covariance matrix is estimated. All securities within a cluster are equally weighted to generate returns to a factor portfolio.

We use these positive-weight factors in a Fama MacBeth (1973) framework in order to compare them to alternative methods of factor identification. In particular, we consider principal components as in Conner and Korajczyk (1988) and pre-specified macro-economic factors as in

Chen, Roll and Ross (1986) as alternatives. We find that our positive-weight factors explain nearly as much out-of-sample cross-sectional variation in stock returns as the PC approach, and considerably more than the CRR approach. When we examine the break-down of these factor portfolios by industrial classifications and size, we learn that certain factors are dominated by certain industries.

Besides using factors derived from the stock universe, we also consider factors derived from the equity mutual fund universe. The intuition behind this is provided by Blake, Elton and Gruber (1996) — namely that the best place to "search" for meaningful factors is on the demand-side for portfolios. Using the mutual fund style classification procedure developed Brown and Goetzmann (1997) we generate non-negative weight portfolios of mutual funds. We find that mutual fund factors are in fact superior to factors estimated in the space of individual security returns. These mutual fund styles are separated in the dimensions of growth vs. income and value vs. glamour.

This paper is organized as follows. The next section describes the methodology and the data. The third section reports the results of our analysis. The fourth section concludes.

II. Methodology and Data

II.1 Stochastic specification

The objective of our analysis is to use past returns to determine a natural grouping of stocks that has some predictive power in explaining the future cross-sectional dispersion in security returns. We begin by classification of securities, with possibly changing factor exposures in an APT economy with K factors, where the factors unknown. That is:

A system of N securities may be expressed as:

$$\begin{aligned} \boldsymbol{R}_t &= \boldsymbol{\alpha}_t + \boldsymbol{\beta}_t' \boldsymbol{f}_t + \varepsilon_t \\ E[\varepsilon_t | \boldsymbol{f}_t] &= \mathbf{0} \\ E[\varepsilon_t' \varepsilon_t] &= \Sigma_t \end{aligned} \qquad \textbf{(2)}$$

$$\begin{aligned} R_{it} &= \alpha_{it} + \beta_{it}' f_t + \varepsilon_{it} \\ E[\varepsilon_{it} | f_t] &= 0 \\ E[\varepsilon_{it}^2] &= \sigma^2_{it} < \infty \end{aligned} \qquad \textbf{(1)}$$

In this context, the selection of factors is equivalent to the identification of portfolios of securities such that they account for all of the common variation in asset returns. This requirement is the motivation for the principal component approach taken by Conner and Korajczyk (1988). Under the assumption that factors are invariant to rotation, the principal component structure of the T by T covariance matrix in the dimension of time yields factors ranked in magnitude of explanatory power with respect to cross-sectional differences in realized returns. It is interesting to note that this procedure in effect chooses factors which maximize the period-by-period R^2 in the Fama MacBeth regressions. One drawback of this approach as Brown (1989) shows, is that, while explanatory power is maximized, interpretations in terms of risk exposures is not. Principal components tends to give the equal-weighted portfolio as the first, and therefore most important factor, even when it is not.

An innovation developed by Lakonishok, Shleifer, and Vishny (1993) and used by Fama and French (1992) for asset pricing, is to form factors by sorting securities into *ex ante* classes. For example, FF group stocks into a highest decile book-to-market category and a lowest decile book-to-

market category. The difference between these is used as a factor. This procedure can be modeled as a classification problem. Suppose the ex post total return in period *t* for any security can be represented as,

$$R_{jt} = \alpha_{Jt} + \beta'_{Jt} I_t + e_{jt} \qquad \textbf{(3)}$$

where security j belongs to category J. β_{Jt} refers to the average exposure of securities in class J to an index capturing the return of that attribute. In the FF framework, stocks are grouped into J = 1...6 according to their size their book to market ratio or their price-earnings multiple. Then factors are formed by equal-weighting the extreme J portfolios and by taking differences. In fact, this method of factor selection does not require that the parameters of this equation be estimated, *ex ante*. Writing the equation as,

$$R_{jt} = \mu_{Jt} + \varepsilon_{jt} \qquad \textbf{(4)}$$

where μ_{Jt} is the expected return for category J conditional upon the factor realization I_t indicates that we need only seek to estimate the mean return of class J at time t in order to approximately estimate a factor based upon this classification.

In the extreme case in which the idiosyncratic return component ε_{jt} has a zero mean ex ante and is uncorrelated across securities, the classification into categories will suffice to explain the cross-sectional dispersion of stock returns to the extent that μ_{Jt} differs across categories. More typically, ε_{jt} will be correlated across securities, and thus the J-class time-series means (or functions of them) may be used as factors in the classical APT equation.

In this paper, we endogenously determine the J-classes, rather than pre-specifying them. We

estimate the groups through an iterative re-location algorithm similar to the K-Means procedure developed by Hartigan (1973) and used by Elton and Gruber (1969). The method can be thought of as a direct estimation of the previous equation, given that the membership of each J-class is unknown. The algorithm requires that the number of groups be pre-specified, although approximate tests exist for determining the number of groups in the data. In addition, no global minimum to the optimization is possible without exhausting all combinations of N choose K. Thus, in practice a local minimum is achieved. Despite these drawbacks, the advantage of this approach is that it allows salient classifications to emerge from the data. We test the efficacy of these classifications using the Fama MacBeth procedure out of sample, as well as using the Chen, Roll and Ross (1986) method to test for positive factor premia. We find that factors formed via our classification algorithm are effective at spreading out-of-sample returns. In addition, they appear to yield positive risk premia over the period of study. Finally, due to the fact that the factors are positive, equal-weight portfolios across all securities in the class, we are able to identify their characteristics in terms of industrial composition, size and fundamental variables.

II.2 Classification Methodology

To identify classes of stocks, we apply the iterative re-location algorithm, SC detailed in Brown and Goetzmann (1997) to the space of returns. For out-of-sample tests, we use 24 months of data meaning that we are effectively classifying securities according to their proximity in 24-dimensional space. We use the number of 1-digit SIC classifications as of a given date as the pre-specified number of clusters, and random securities as initial "centers" to the algorithm. The SC procedure seeks exchanges of securities among groups so as to minimize within-group sums of

squared deviations from the group centers. It uses the Euclidean distance among observations in the manner of Hartigan's K-Means algorithm. [1] It is important to stress how this criterion differs from factor analysis methods typically used in empirical APT estimation. Observations are not de-meaned — classification depends explicitly on the drift as well as the variation about it. Proximity is not defined in terms of summed products of de-meaned observations (i.e. covariance) but in terms of summed squared differences. Given this unusual criterion, one might expect Finance theory to provide little guidance regarding what to expect of the resulting groups, however there are, if fact, some useful interpretations.

Suppose that we applied the SC algorithm to a set of Arrow-Debreu securities and this set included redundant securities. Further assume that the dimension of time proxies for the dimension of states. The algorithm would classify all redundant securities correctly — all securities with a unit payoff (or less) in a given state would end up in the same group. With incomplete markets, in which no pure Arrow-Debreu securities exist, the algorithm groups securities according to similarity in payoff structures. It forces together all securities with closely matched payoff returns by state, and it forces apart securities with widely divergent patterns of state payoffs. In fact, it chooses a parsimonious set of maximally spanning portfolios, given the constraint that portfolio weights are positive and equal across within-group assets, and that all assets must belong to one of the portfolios.[2] Another interpretation of the security clusters, in light of equation 4 is that all securities are interpretable as noisy manifestations of a parsimonious set of fundamental portfolios, and the algorithm simply sorts securities into the appropriate portfolios according to a maximum-likelihood criterion. This analysis suggests that the longer the time period, the broader the range of realized states, and consequently, the more accurate the estimate of spanning portfolios. By the same token,

however, a few periods should be sufficient to provide some information for classification as long as the number of states exceeds the number of pre-specified groups.

We apply the SC algorithm in two contexts. For out-of-sample Fama MacBeth tests, we use monthly CRSP data from 1976 through 1992. Each period, we estimate clusters with twenty-four months of individual security data and then perform cross-sectional regressions on the following year returns. We also estimate clusters over the entire 156 months of data, conditioning upon survival the entire period, and then estimate premia for the factors.

III. Cross-Section Tests

III.1 Cross-sectional regressions on classification codes

III.1.a Dummy variable regressions

As a preliminary to formal pricing tests, we examine whether the classifications obtained by the SC algorithm explain out-of-sample cross-sectional differences in returns. This can be though of as a Fama MacBeth style procedure, in which the regressors each period are dummy variables, rather than factor loadings. For each 24-month period in our sample from 1976 through 1992, we estimate SC classifications of the CRSP monthly return data (2,600 stocks) conditional upon the number of major SIC groups that exist at the beginning of that period (eight). For comparison, following Brown and Goetzmann (1997), we also use as regressors clusters formed (1) in the space of principal component factor loadings, (2) in the space of Sharpe coefficients, and (3) on major SIC classifications.[3] Summary statistics across all 14 year periods for these results are reported in the first panel of Table 1. Note that the SC classification works best of the four classification schemes.

While we cannot reject the hypothesis that the mean and median R^2 is equal for the top performers, groups formed in the space of PC weights and in the space of returns are both marginally more informative than major SIC codes at explaining cross-sectional differences out-of-sample. Clusters formed in the space of returns perform slightly better than clusters formed in the space of PC loadings.

III.1.b Factor-loading regressions

Next, we perform Fama-MacBeth cross-sectional regressions on factor loadings themselves. We take the SC centers as factors and estimate loadings for each security on these factors. We compare and finally Chen-Roll-Ross - style regressors, on passive asset indices. Surpisingly , SC centers work the best, on average. This is most surprising in light of the fact that principal components are chosen to maximally spread returns in sample. The best median performer is the factor loadings. In general, use of factor loadings, as opposed to classifications increases explanatory power only a little. In the current formulation, we are not using any correction (see Shanken, 1992) for the errors-in-variables bias in the factor coefficients. In the next section, we seek to reduce the errors in variables problem through forming industry portfolios.

IV. Risk Premia Tests

IV.1 Comparing SC premia to PC and macro-series

Following Chen, Roll and Ross (1986), we estimate clusters over the whole time period (228 months) and compare the cross-section effects to two other approaches. First, seven clusters are

formed using 228 months of stocks surviving the whole period. Second, principal components are estimated using same data. Third, we form portfolios based upon 2 digit SIC cides (56 groups). These are used in place of the size portfolios in CRR. The motivation for the use of portfolios is the classic errors-in-variables problem in the Fama-MacBeth test, i.e. factor loadings in the first pass are estimated with error. When used as regressors in the second pass, they no longer generate consistent estimates of risk premia. Aggregation into portfolios reduces the problem, but does not eliminate it. We formed industry portfolios because it seemed likely that loadings on factors would be similar within industries.

Next, we estimate loadings over the whole period on three sets of regressors: the seven cluster centers (i.e. our "cluster" factors), the first seven principal components (estimated from whole set of stocks, not the industry indices), and finally, loadings on us capital market indices (small stocks, S&P, High-yield bonds, Corporate bonds, Government bonds, Commercial paper and, Treasury-bills). Finally, each month we perform a cross-sectional regression to get the premium. We also save the R^2 for each month.

The mean and median of the R^2 series from each method suggests that the SC procedure generates factors that work considerably better in-sample than loadings on exogenously specified financial indices. The mean and median R^2 for the SC centers are 0.195 and 0.175, while the R^2 on the cap. mkt. indices are 0.169 and 0.145 respectively. Since this is an in-sample test, the cross-sectional regressions on the principal component factor loadings provide a maximum R^2 measure, with a mean and median of .212 and 0.208, respectively. Note that clusters work much better than the capital market indices. They do almost as well as the principal components on average which, after all, are constructed to explain variation.

Following CRR, we estimate the time-series of factor premia, and test whether they are different from zero. Table 2 summarizes the risk-premia time-series. Notice that the principal component factors do not have positive premia, despite spreading returns. This is because they are derived solely from the covariance matrix, and contain both positive and negative weights and they do not sum to one. We have not re-constructed matching portfolios as in Conner and Korajczyk.

IV.2 Comparing SC premia to premia derived from the mutual fund universe

A recent paper by Blake, Elton and Gruber (1996) proposes a novel idea — that the best place to search for pricing factors is on the demand-side, rather than the supply-side. If factors indeed reflect sources of risk that investors care about, we might expect to find them as salient determinants of fund returns. Mutual funds are portfolios loaded on particular factors that have differing positive expected returns. There are other attractive motivations for deriving factors from the mutual fund universe. First, funds are diversified, and this reduces estimation error. Second, funds pursue dynamic strategies. For example, a value manager may buy stocks with low PE ratios and sell these stocks when the PE ratio increases. This activity is analogous to the Fama and French (1992) sorting each year. This is not possible with principal components analysis, or with clustering applied to the entire time-series matrix of stock returns over 228 months. Portfolio weights are not time-varying.[4]

Using monthly data from Morningstar, Inc. all-equity funds over the period equal to our study, we estimated eight clusters using a variation on the SC algorithm which accounts for heteroskedasticity. The method is fully described in Brown and Goetzmann (1997). The advantage of using these styles is that we have a clear idea of their relationship to fund manager strategies. The

styles include an equity income group, a growth and income group, a growth group, a value group, a glamour group, an international group, a global timing group and a metals group — all of which emerge endogenously from the application of the style classification algorithm. We take the mutual fund styles identified in the Morningstar data, and define the style portfolios as factors for pricing of stocks.[5] We find that the mutual fund style centers work extraordinarily well at explaining cross-sectional variation in stock returns. The mean and median adjusted R^2 are .210 and .179 respectively.[6] While a test for differences is significant, these values exceed the results based upon clustering the space of stock returns themselves. Of more interest are the time-series of risk premia generated for the mutual fund style factors. Note that five of the eight factors yield t-statistics over two, as opposed to two each for SC factors and two for capital market portfolios. This higher level of reliability is evidently due to higher mean values, rather than lower variability. Most of the premia associated with the fund style factors are about 1% per month.

V. Factor Interpretation

V.1 Cross-tabulation

One way to interpret the cluster-based factors is to examine the composition of the portfolios. The CRSP data contains SIC and capitalization information that we use for cross-tabulation with the clusters formed over the entire period. Table 3 reports the industrial composition of the groups, and table 4 reports the capitalization of the groups. Factor 6 is clearly a market factor, with positive weights in many sectors of the economy. Factor 1 is a utility industry factor, 2 comprises mostly small cap stocks, and factor 3 is a mining and minerals factor. Factor 4 appears to be a high cap

energy factor. The remaining factors appear to be difficult to classify, although factor 7 appears to be concentrated in the services sector.

VI. Conclusion

The motivation for using alternate methods and data sets for deriving factors is the desire to understand what risk factors are perceived and priced by equity investors. While statistical procedures such as principal component analysis provide explanatory power, they provide little intuition. The pre-specified factor structure identified by CRR and FF provide intuition regarding priced factors, but are given exogenously. Consequently the possibility of identifying alternative structures endogenously, or from other sources as in Blake, Elton and Gruber (1996) holds potential for powerful as well as interpretable factors. In this paper, we apply simple an intuitively appealing criterion for identifying structures in security returns. Clustering methods are consistent with a finite state-space representation of security returns in which a parsimonious set of portfolio factors are desired. These yield factors which spread returns out of sample. In the classic CRR approach to time-series risk premia estimation, several cluster-based factors are associated with significantly positive risk premia. These clusters, in turn are related to specific industry groups.

Turning to the space of mutual fund returns, we find that factors derived from equity fund styles perform better than factors derived from the stocks themselves.

References

Blake, Christopher, Edwin Elton and Martin Gruber, 1996, Factors in security returns (??), Working Paper, Stern School of Business, NYU, 1996.

Brown, Stephen J., 1989, The number of factors in security returns, Journal of Finance 44:5, 1247-1262.

Brown, Stephen J. and William N. Goetzmann, 1997, Mutual fund styles, Journal of Financial Economics, Forthcoming.

Chen, Naifu, Richard Roll and Stephen Ross, 1986, Economic forces and the stock market, Journal of Business, 59, 383-403.

Conner, Gregory and Robert Korajczyk, 1988, Risk and return in an equilibrium APT: Application of a new test methodology, Journal of Financial Economics, 21, 255-290.

Elton, Edwin and Martin Gruber, 1970, Improved forecasting through the design of homogeneous groupings, Journal of Business 44, 432-450.

Elton, Edwin, Martin Gruber, Sanjiv Das and M. Hlavka,1993, Efficiency with costly information: A reinterpretation of evidence for managed portfolios, Review of Financial Studies 6, 1-22.

Fama, Eugene, J. MacBeth ,1973, Risk, return and equilibrium: Empirical tests, Journal of Political Economy, 71, 67-636.

Fama Eugene and Kenneth French, 1992, The cross-section of expected stock returns, Journal of Finance, 47, 427-465.

Grinblatt, Mark and Sheridan Titman, 1985, Factor pricing in a finite economy, Journal of Financial Economics, 12, 497-507.

Hartigan, John A., 1975, Clustering Algorithms (John Wiley and Sons, New York, NY).

Lakonishok, Josef, Andrei Shleifer, and Robert Vishny, 1993, Contrarian investment, extrapolation and investment risk, Journal of Finance 49, 1541-1578.

Lehmann, Bruce N. and David Modest, 1988, The empirical foundations of the arbitrage pricing theory, Journal of Financial Economics, 21, 213-254.

Roll, Richard and Stephen Ross, 1980, An empirical examination of the arbitrage pricing theory, Journal of Finance, 35, 1073-1103.

Shanken, Jay, 1992, On the estimation of beta pricing models, Review of Financial Studies, 5,1-34.

Table 1: Cross-Sectional Return Variance Explained by Ex-Ante Classification Methods and Factor Loadings

	Regressing Returns on Classifications: Adjusted R^2				Regressing Returns on Factor Loadings: Adjusted R^2			
Test period	Return based classifica-tions (GSC procedure)	Classific a-tions based on Sharpe coefficie nts	Classifica -tions based on principal components	Classific a-tions based on 1-digit SIC codes	Constraine d pre-specified factors (Sharpe procedure)	Principal factor loadings	GSC centers	Unconstra ined pre-specified factor loadings
mean	0.061	0.019	0.058	0.051	0.038	0.066	0.069	0.053
median	0.061	0.017	0.059	0.051	0.034	0.064	0.059	0.045
std. deviation	0.025	0.013	0.022	0.027	0.027	0.023	0.024	0.026

This table uses CRSP total return data over the period 1976 through 1994 in a Fama-MacBeth procedure applied to non-overlapping annual test periods. A 24-month estimation period to estimate classifications and factor loadings, followed by a 12 month period in which annual returns are calculated and used in cross-section regressions. The number of classifications in each period is specified by the number of single-digit SIC codes. The cross section of test period returns on funds are regressed against (K - 1) dummy variables, where $\delta_{ki} = 1$ for fund I in category k and zero otherwise. The first column gives adjusted R^2 for the categories given by the GSC procedure described in the text, the second and third columns correspond to categories based on constrained Sharpe coefficients and principal factors procedures (using the SC procedure), while the fourth column uses the Weisenberger style categories (1978-93) and Morningstar categories (1994). These are compared to the adjusted R^2 obtained by using the Sharpe coefficients (Column 5), factor loadings (Column 6), loadings on the SC style centers (Column 7), and the unconstrained loadings on capital market returns used to estimate the Sharpe coefficients (Column 8). The data are total returns to U.S. equity mutual funds, excluding sector funds, but including international funds, over the period 1976 through 1994.

Table 2
Risk premia derived from SC clusters

	t-stat	prob.	monthly mean	monthly std
cluster 1	1.874	0.062	0.009	0.071
cluster 2	1.969	0.050	0.013	0.097
cluster 3	0.762	0.447	0.007	0.132
cluster 4	1.572	0.117	0.011	0.101
cluster 5	0.487	0.627	0.007	0.211
cluster 6	2.133	0.034	0.013	0.095
cluster 7	2.517	0.013	0.018	0.108

Risk premia derived from cap. market loadings:

	t-stat	prob.	monthly mean	monthly std
small	2.382	0.018	1.380	8.745
sp	2.051	0.041	1.019	7.499
hiyld	1.749	0.082	0.700	6.042
ltc	1.750	0.081	0.746	6.439
ltg	1.820	0.070	0.928	7.700
cp	1.117	0.265	0.048	0.648
tb	1.089	0.277	0.043	0.599

Risk premia derived from principal component loadings:

	t-stat	prob.	monthly mean	monthly std
comp 1	1.455	0.147	0.007	0.073
comp 2	-1.418	0.158	-0.007	0.070
comp 3	1.238	0.217	0.006	0.075
comp 4	0.181	0.857	0.001	0.089
comp 5	1.238	0.217	0.007	0.082
comp 6	-1.491	0.137	-0.008	0.078
comp 7	-0.814	0.416	-0.007	0.137

Risk premia derived from mutual fund styles:

	t-stat	prob.	monthly mean	monthly std
growth and income	2.089	0.038	0.010	0.069
growth	2.294	0.023	0.011	0.075
income	2.092	0.038	0.008	0.057
value	2.124	0.035	0.011	0.076
global timing	1.879	0.062	0.010	0.077
glamour	2.743	0.007	0.015	0.085
international	1.726	0.086	0.013	0.111
metal funds	0.455	0.650	0.004	0.130

Table 3: Number of Firms by Two Digit SIC Code and Return Based SC Classification

	1	2	3	4	5	6	7	**Sum**
Agricultural Production--Crops		1					1	**2**
Apparel And Accessory Stores						2		**2**
Apparel And Other Textile Products		7				5	4	**16**
Auto Repair, Services, And Parking		2				1		**3**
Automotive Dealers & Service Stations						1		**1**
Building Materials & Garden Supplies		2						**2**
Business Services		4				6	1	**11**
Chemicals And Allied Products	3	7		4	2	41	2	**59**
Coal Mining			1	3		1		**5**
Communication	4					3		**7**
Depository Institutions						10	1	**11**
Eating And Drinking Places		1				1	1	**3**
Educational Services						1	1	**2**
Electric, Gas, And Sanitary Services	100			8		3		**111**
Electronic & Other Electric Equipment		7				20	22	**49**
Engineering & Management Services		2				1	2	**5**
Fabricated Metal Products		9				12	11	**32**
Food And Kindred Products	11	4				12	1	**28**
Food Stores						5		**5**
Furniture And Homefurnishings Stores						1	2	**3**
Furniture And Fixtures						1		**1**
General Building Contractors					1	1		**2**
General Merchandise Stores		1				9	3	**13**
Health Services						2	2	**4**
Heavy Construction, Ex. Building		4		1				**5**
Holding And Other Investment Offices	29	17	1	7		20	4	**78**
Hotels And Other Lodging Places		1				2	2	**5**
Industrial Machinery And Equipment	1	16		2	1	23	11	**54**
Instruments And Related Products		2				12	6	**20**
Insurance Agents, Brokers, & Service	1							**1**
Insurance Carriers		1				8		**9**
Leather And Leather Products		3				3	2	**8**
Local And Interurban Passenger Transit		1				2	1	**4**
Lumber And Wood Products		1				3	2	**6**
Metal Mining		2	9	1		2		**14**
Miscellaneous Manufacturing Industries		3				2	1	**6**
Miscellaneous Retail		1				6		**7**
Motion Pictures						2		**2**
Nondepository Institutions	1	3				5	1	**10**
Nonmetallic Minerals, Except Fuels		3						**3**
Oil And Gas Extraction	2		1	24			1	**28**
Paper And Allied Products		1				14	1	**16**
Personal Services						3	1	**4**
Petroleum And Coal Products	1	1		15		1		**18**
Primary Metal Industries		5	3			13	2	**23**
Printing And Publishing		4				16		**20**
Railroad Transportation		2				2		**4**
Real Estate	1	3					4	**8**
Rubber And Misc. Plastics Products		3				8	3	**14**
Security And Commodity Brokers							2	**2**
Services, Nec						2	1	**3**
Special Trade Contractors							1	**1**
Stone, Clay, And Glass Products		4				5	2	**11**
Textile Mill Products		2				1	3	**6**
Tobacco Products	1	1				3		**5**
Transportation By Air		1				4	3	**8**
Transportation Equipment		8				16	9	**33**
Transportation Services							1	**1**
Water Transportation				1		2		**3**

Wholesale Trade--Durable Goods			3			4	3	**10**
Wholesale Trade--Nondurable Goods			1			6	1	**8**
	Sum	**155**	**144**	**15**	**66**	**4**	**329**	**122**	**835**

Table 4: Average Capitalization ($000) by Two Digit SIC Code and Return Based SC Classification

	1	2	3	4	5	6	7	**Average**
Agricultural Production--Crops		$21,318					$6,946	**$14,132**
Apparel And Accessory Stores						$339,061		**$339,061**
Apparel And Other Textile Products		$12,244				$76,470	$50,446	**$41,865**
Auto Repair, Services, And Parking		$18,312				$121,619		**$52,748**
Automotive Dealers & Service Stations						$16,152		**$16,152**
Building Materials & Garden Supplies		$135,045						**$135,045**
Business Services		$11,204				$250,048	$56,952	**$145,641**
Chemicals And Allied Products	$4,852,820	$41,357		$394,129	$7,170	$1,328,716	$78,938	**$1,204,645**
Coal Mining			$11,878	$402,049		$55,054		**$254,616**
Communication	$7,358,792					$633,276		**$4,476,428**
Depository Institutions						$304,011	$107,174	**$286,117**
Eating And Drinking Places		$11,681				$2,342,939	$15,993	**$790,204**
Educational Services						$22,627	$2,421	**$12,524**
Electric, Gas, And Sanitary Services	$378,213			$582,704		$154,929		**$386,917**
Electronic & Other Electric Equipment		$24,967				$1,081,646	$74,645	**$478,569**
Engineering & Management Services		$5,628				$14,835	$5,899	**$7,578**
Fabricated Metal Products		$26,866				$287,948	$45,794	**$131,278**
Food And Kindred Products	$1,351,595	$20,528				$353,014	$56,574	**$687,228**
Food Stores						$163,221		**$163,221**
Furniture And Homefurnishings Stores						$61,977	$237,761	**$179,166**
Furniture And Fixtures						$624,146		**$624,146**
General Building Contractors					$49,703	$115,616		**$82,659**
General Merchandise Stores		$30,137				$2,182,128	$18,609	**$1,517,317**
Health Services						$34,456	$12,068	**$23,262**
Heavy Construction, Ex. Building		$39,790		$538,266				**$139,485**
Holding And Other Investment Offices	$275,252	$32,134	$283,200	$1,223,260		$381,750	$24,139	**$321,874**
Hotels And Other Lodging Places		$64,906				$386,627	$12,876	**$172,782**
Industrial Machinery And Equipment	$686,807	$303,505		$133,816	$5,587	$1,996,786	$89,974	**$976,517**
Instruments And Related Products		$21,633				$2,659,026	$16,533	**$1,602,539**
Insurance Agents, Brokers, & Service	$788,122							**$788,122**
Insurance Carriers		$64,096				$578,348		**$521,209**
Leather And Leather Products		$7,192				$202,190	$72,203	**$96,569**
Local And Interurban Passenger Transit		$25,301				$388,130	$10,320	**$202,970**
Lumber And Wood Products		$6,521				$1,106,844	$27,154	**$563,560**
Metal Mining		$202,575	$75,959	$809,894		$556,055		**$215,056**
Miscellaneous Manufacturing Industries		$29,980				$61,418	$6,814	**$36,599**
Miscellaneous Retail		$7,942				$134,099		**$116,077**
Motion Pictures						$817,196		**$817,196**
Nondepository Institutions	$115,194	$23,085				$319,209	$8,243	**$178,874**
Nonmetallic Minerals, Except Fuels		$176,534						**$176,534**
Oil And Gas Extraction	$173,670		$564,857	$609,760			$3,562	**$555,357**
Paper And Allied Products		$51,509				$1,296,251	$54,308	**$1,140,833**
Personal Services						$108,482	$14,280	**$84,932**
Petroleum And Coal Products	$19,855,150	$15,494		$1,911,898		$235,767		**$2,710,271**
Primary Metal Industries		$61,408	$382,433			$768,422	$28,752	**$500,058**
Printing And Publishing		$20,032				$228,516		**$186,819**
Railroad Transportation		$30,906				$664,635		**$347,770**
Real Estate	$24,562	$23,838					$6,975	**$15,497**
Rubber And Misc. Plastics Products		$135,992				$329,523	$8,214	**$219,200**
Security And Commodity Brokers							$269,424	**$269,424**
Services, Nec						$132,525	$8,096	**$91,048**
Special Trade Contractors							$8,060	**$8,060**
Stone, Clay, And Glass Products		$28,303				$451,833	$38,421	**$222,656**
Textile Mill Products		$32,250				$94,199	$22,012	**$37,456**
Tobacco Products	$988,491	$32,409				$1,194,484		**$920,870**
Transportation By Air		$85,388				$425,703	$26,912	**$233,617**

Transportation Equipment		$108,501				$1,698,559	$57,295	**$865,473**
Transportation Services							$36,367	**$36,367**
Water Transportation				$113,088		$226,493		**$188,691**
Wholesale Trade--Durable Goods		$14,293				$273,096	$6,421	**$115,453**
Wholesale Trade--Nondurable Goods		$23,724				$62,648	$15,834	**$51,931**
Average	**$822,390**	**$72,323**	**$179,391**	**$924,980**	**$17,407**	**$880,948**	**$51,657**	**$596,202**

Notes

1.Although Brown and Goetzmann (1997) show that it is useful in many cases to scale distances by variance, and other distance metrics, namely Mahalinobis measures have also been proposed as preferable in the presence of heteroskedasticity and serial correlation, this greatly increases the computational complexity.

2. This suggests an interesting extension of the algorithm. It may be possible to relax the requirement that factor portfolios exhaust securities, and seek instead to identify a parsimonious set of positive-weight portfolios that span payoffs via a similar algorithm.
For example, maintaining the requirement of equal-weighting, but relaxing the exhaustion constraint simply creates a sub-set of membership to a group in the algorithm — one sub-set of securities within the group are used for determining the group center, while the broader set are used in the calculation of within-group sums of squares. While relaxation of the equal-weight constriant is more difficult, it may be possible as well.

3. We cluster in the space of "Sharpe coefficients" (for details, see Sharpe, 1992), which are estimated on a set of eight capital market indexes obtained from Ibbotson Associates. These include an IPO index, S&P, small stocks, high-yield bonds, corporate bonds, government bonds, t-bills and commercial. These are estimated via a constrained optimization procedure under the assumptions that the weights remain fixed over the estimation period, that they are nonnegative, and that they sum to one. The weights may thus be interpreted as portfolio weights for passive, investable indexes. This method is used because of its ease of factor interpretation.

4. Note that in our out-of-sample test, weights are allowed to vary through time. We only constrain them to be constant over 24-month intervals.

5.Blake, Elton and Gruber (1996) do something analogous to this. In their quest for a "fifth" stock factor, they find that a portfolio based upon growth funds explains variation in stock returns better under certain criteria that principal components derived from the stock universe.

6. These are the statistical summaries of the time-series of R^2 for each monthly cross-sectional regression.

	mean	median
cluster	0.19450	0.17545
cap. mkt. indices	0.16933	0.14531
pr. comps	0.21207	0.20761
mutual fund styles	0.21032	0.17938