Some Aspects of the Optimal Grouping of Stocks

by

Geoffrey J. Lauprete

B.S., Tufts University (1993)

Submitted to the Sloan School of Management
in partial fulfillment of the requirements for the degree of

Master of Science in Operations Research

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 1998

Signature of Author ....
Sloan School of Management
14 November 1997

Certified by ....
Roy E. Welsch
Professor of Statistics and Management Science
Thesis Supervisor

Accepted by ....
Robert Freund
Co-Director, Operations Research Center

Some Aspects of the Optimal Grouping of Stocks

by

Geoffrey J. Lauprete

Submitted to the Sloan School of Management
on 14 November 1997, in partial fulfillment of the
requirements for the degree of
Master of Science in Operations Research

Abstract

This thesis develops a probabilistic model to characterize cross-sectional stock return behavior using industry classification information. Stock returns are modeled as normally distributed, with mean and variance for a given time period being parametric functions of firm-specific variables and industry affiliation. Then, clustering algorithms are presented, in an attempt to optimize the assignment of stocks to industry sectors. The algorithms are initialized with a given industry classification, and work by merging sectors which are "closest" according to some distance metric. The optimality of an industry classification is measured in terms of the Akaike Information Criterion (AIC) of the model which uses that classification.

Thesis Supervisor: Roy E. Welsch
Title: Professor of Statistics and Management Science

Contents

List of Tables

List of Figures

Acknowledgements

I would like to thank Jim White from TASC for guidance and intellectual support, and TASC for financial support. Support was also provided by the Finance Research Center and the Center for Computational Research in Economics and Management Science, both at MIT, and NSF Grant DMS-9626348 (P.I.: A. Samarov). Thanks are due to Prof. Roy Welsch, Alexander Samarov, and Peter Kempthorne for valuable discussions and shared insight. Finally, I would like to thank MIT and the ORC for providing an environment in which I could pursue my academic and research interests as they naturally arose.

Chapter 1

Introduction

Industry classifications of stocks are widely available, but vary between the financial - or economic - institutions which provide them. One example of industry classification is the coding system provided by the U.S. government Office of Management and Budget, the Standard Industrial Classification (SIC). Its coding of industries is composed of 4 digits. The first two digits represent a broad industrial class, such as Construction or Transportation. The last two digits provide further subdivisions into more precisely defined industry sectors. Another example of industry classification is that provided by the financial services company Vestek, which uses a total of 69 distinct sectors - the later classification is used in this thesis. Industry sectors are typically constructed to form mutually exclusive and collectively exhaustive sets of stocks within a given population. Hence, the industry sectors form a partition of the population of stocks. Even though such classifications typically assign each stock to one industry sector, they differ in their definitions of industry sectors and in their judgement of what stocks should be put in the same sector.

In this thesis, we show how the information contained in such industry classifications can help to model stock returns. Specifically, we develop parametric probabilistic models which describe the cross-sectional behavior of returns, conditional on industry classification information and other relevant explanatory variables. Most importantly, industry classification information allows us to model the inherent heteroskedasticity of cross-sectional returns. Our reasoning is that returns in different industries will have different cross-sectional variances. The fact that we are modeling the heteroskedasticity of returns should give us more efficient parameter estimates. Since we are ultimately interested in the values of the estimated parameters and their stability over time, our methodology allows us to have greater confidence in the meaning of our results.

Furthermore, the models which we develop allow us to rank industry classification schemes. We reason in terms of the models to decide which industry classifications are preferred. We examine the Akaike Information Criterion (AIC) - a measure of fit related to the predictive ability of the model, see below for details on the AIC - of our models under different classification schemes, to decide what classifications serve the pupose of our modeling best. Simply put, a classification is preferred if it leads to a more optimal AIC value. Given an initial industry classification, we then show how a more optimal classification may be obtained, by merging sectors by means of clustering algorithms.

A word is in order about how this research fits into the current large body of work concerning cross-sectional stock returns. The focus of a number of papers in the last decade has been to determine what variables seem to affect returns in cross-sectional OLS regressions of stock returns on firm-specific explanatory variables. A consensus seems to have emerged over the fact that market equity (ME) of a company - market equity is defined as the number of outstanding shares multiplied by the price per share - is negatively correlated with its returns. Other variables which have captured the attention of researchers, and which have been shown to be significant on certain datasets are earnings-to-price (E/P), and book-to-market ratios (BE/ME). Fama and French (1992) show that indeed the above variables are significant determinants of return on monthly cross-sections from the U.S. stock market - CRSP datasets. Furthermore, they show that β is not a significant discriminant of cross-sectional returns when other firm-specific variables are included in OLS regressions. Our research can be seen as taking for granted the fact that a selection of variables - namely ME, E/P, and BE/ME - may be significant determinants of return. We then attempt to obtain more efficient estimates of the parameters affecting these variables, by modeling the inherent heteroskedasticity of returns between industry sectors. That is, some sectors are defined in such a way as to have a large variations of return for any given month, whereas other sectors have returns that are more tightly distributed around their mean. χ^2 tests support the hypothesis of heteroskedasticity between industry sectors. Furthermore, we conclude that indeed we obtain more efficient estimates of the parameters, since our more general model, where we have explicitely taken into account the heteroskedasticity of returns, has higher AIC values on our datasets.

The organization of this thesis is as follows. Chapter 2 states the goals of our research and mentions application areas for our results. Chapter 3 reviews the literature which deals with factor models of stock returns - models with explain returns in terms of explanatory variables. We also refer to studies which have considered the problem of grouping stocks, using methods such as factor analysis or clustering algorithms. Chapter 4 presents the data we use. Chapter 5 supplies the details of the parametric probabilistic models of returns that we consider. We first consider the most general specification of our model. We then consider certain restrictions on the parameters of the general model, which imply respectively OLS regression, and the groupwise heteroskedastic (GWH) assumption. In chapter 6 we apply our models to selected datasets. Chapter 7 presents the clustering algorithms, and mainly the K-means algorithm, which we use to group industry sectors. Chapter 8 discusses the results of the clustering algorithms In chapter 9 we provide a summary of the results presented in this paper, and we present possible directions for future research.

Chapter 2

Motivation

The goals of this research are twofold:

1. to specify the form of probabilistic models which describe the cross-sectional distribution of stock returns. That distribution should be a function of the available information, namely firm-specific variables, and industry affiliation. We focus our attention on cross-sectional models, and will not make any assumptions about the model parameters' evolution over time. We will therefore estimate the models at each time period. We will, however, model the cross-sectional heteroskedasticity of returns.

2. to examine and evaluate methods to cluster sectors into groups in order to reduce the dimensionality of the industry classification. The intuition behind this approach is that too many sectors may lead to models that overparametrize the actual distribution of returns. We expect that the models which use a more concise industry classification will optimize our model selection criterion, the AIC. Hence, finding the "best" industry classification would be equivalent to choosing the classification that optimizes the AIC of the model. We expect that clustering algorithms will allow us to obtain more optimal industry classifications, starting from an initial industry classification, such as the one from VESTEK.

2.1 Statement of Problem

2.1.1 Parametric Probabilistic Model of Returns

Returns are assumed to be random variables having a normal distribution with mean and variance being linear functions of firm-specific explanatory variables. The parameters of these firm-specific explanatory variables may depend on the industry sector to which the stock belongs, in which case industry classification affects the form of the model. This model specification is general, in that our model can use any explanatory variables which seem relevant in explaining returns. We provide some insight on which variables are appropriate, but this is not the central focus of our study. We take as given the conclusion of Fama and French (1992) that three firm-specific variables - ME, E/P, and BE/ME - determine the parameters of cross-sectional stock return distributions. Also, we deal exclusively with a cross-section of returns, and leave the analysis of the time-series aspect of returns for future research. [1]

We rank models on the basis of their Akaike Information Criterion (AIC), which is defined as

$$-2(\text{Loglikelihood of Model}) + 2(\text{Number of Parameters in Model})$$

Remember - or notice - that lower values of the AIC are indicative of better model fit. The AIC is a widely used model selection criterion. The AIC of a model on a given dataset, where the parameters have been estimated by Maximum Likelihood. provides an unbiased estimate of the loglikelihood of the model on a future dataset[2]. The properties of the AIC depend, of course, on the assumption that the future data come from the same generating process as produced the initial data. In practice the AIC is useful because it provides a penalty for the number of parameters used in estimating the model. It is common knowledge that using too many parameters is bad for prediction, or for using the model out of sample - i.e. on another dataset. Hence, the AIC controls for overfitting by penalizing models that use too many parameters.

[1] Hence, no matter what our disposition is concerning the research of Fama and French (1992), we would be limited to using firm-specific variables and sector dummy variables. Indeed, macroeconomic variables, such as inflation or unemployment, which could potentially explain the time-series behavior of returns, have no role to play when studying a single cross-section of returns at a time, as we do here - they do not vary across observations of stock returns.

[2] See for example Akaike (1973).

We further justify the use of the AIC by calculating the likelihood of the model on several datasets besides the one which is used to estimate the parameters. We verify that the AIC does allow us to choose models which yield a higher likelihood on other datasets.

2.1.2 Clustering Algorithms

The second goal of this research is to find better industry classifications of a population of stocks, starting with an initial industry classification. An industry classification as used here means the partition of the population of stocks into sets. "Better" is defined in terms of the AIC of the models that use industry classification information. We expect that by clustering sectors that are close enough according to some distance metric to be defined, we will have decreased the dimensionality of the industry classification, which in turn should allow us to obtain models with more optimal AICs. Notice that the first and second goals are complementary, since the clustering algorithm needs the probabilistic model to function, and is intended to yield a more parsimonious model with better predictive power, reinforcing the first goal.

The industry classifications that we consider are subsets of the power set of an arbitrary initial industry classification, such as one given by financial or economic institutions. This means that we consider merging sectors from the original industry classification, but not splittings of those sectors. The idea is to limit the dimension of the space through which we search for a better classification. Indeed, enumeration of all the different partitions of the population of stocks is not feasible, given time and computational constraints. To see why, consider that the number of ways of grouping 2000 stocks - this is the average number of stocks in any one of our datasets - into 2 sectors is

$$(1/2!)\Sigma_{j=0}^{2}(-1)^{2-j}\begin{pmatrix} 2 \\ j \end{pmatrix} j^{2000} \simeq 2^{2000}$$

In general, the number of ways of sorting n objects into k nomempty groups is given by the expression [3]

$$(1/k!)\Sigma_{j=0}^{k}(-1)^{k-j}\begin{pmatrix} k \\ j \end{pmatrix} j^{n}$$

[3] King (1966)

Adding these numbers for $k = 1,2, \ldots , n$, we obtain the total number of ways of sorting n objects into groups.

To deal with the explosive combinatorial nature of our problem, we choose to develop a clustering algorithm which takes as given an initial, financial industry classification, and determines which mergings of industry sectors, according to some appropriate distance metric between sectors. Ideally, we would have liked to navigate through the entire space of sector mergings and splittings in an AIC-optimizing direction - a difficult task -, but short of this lofty goal, we can arrange to try merging a restricted set of sectors, and keep track of the AIC. We can then choose which combinations of sectors from our restricted choice set results in the highest AIC. Hence, since we cannot determine the best industry classification from the total set of all possible partitions of the population of stocks, we have to settle for a sub-optimal, but, hopefully, near optimal solution.

2.2 Application Areas for Probabilistic Models of Stock Returns and Clustering Algorithms for Industry Sectors

Both the stock return model and the clustering algorithm are of interest in any application that requires estimates of expected stock returns. Indeed. in a probabilistic framework such as we consider, stock returns are assumed to have a mean which depends on explanatory variables and industry classification. Once the model is constructed, expected returns can be used for [4]

1. **selecting portfolios**. The estimated parameters of the model and historical average of the corresponding portfolio variables can be used to estimate the expected return on the portfolio.

2. **estimating the cost of capital**. The expected return is the cost of capital. This cost can be used as a discount factor when evaluating future income. to come up with an estimate of the firm's present value.

Also, the manner in which we optimize the choice of the industry classification means that we are choosing to group stocks whose behavior is statistically unique and different from the

[4] Fama and French (1993)

rest. This optimized industry classification can then be used for the design of index numbers, which are the average stock return over industry sectors [5]. An index for a particular industry should be highly correlated with the factors affecting that industry, and uncorrelated with other factors. Our model attempts to reveal which sectors have a similar behavior and can therefore be merged, and which sectors are unique and should be left alone - i.e. not merged.

Finally, our models attempt to give us the most complete specification of the relation between stock returns and the explanatory variables included in the models. Such models allow us to estimate how sensitive returns are to the different variables, and using a more complete information set, that includes industry classification information should increase our confidence in our estimates. This is true if, as we are about to do, we model the inherent heteroskedasticity of returns within sectors. Incorporating this knowledge of heteroskedasticity will give us more efficient parameter estimates, where efficiency is used the classical statistical sense of the word, meaning lower variance[6].

The confidence that we have in being able to accomplish the above tasks depends on the confidence which we have in the underlying model. For this reason, we monitor the AIC, a performance measure, and choose the model which optimizes it, among all model specifications we consider.

[5] King (1966)

[6] see for example Greene (1993), p.387.

Chapter 3

Literature Review

3.1 Probabilistic Factor Models of Stock Returns

In this section we review models that attempt to explain the probabilistic nature of stock returns, conditional on explanatory variables - hence the name "factor models".

3.1.1 CAPM and APT: Expected Returns

The first model we look at, the Capital Asset Pricing Model (CAPM), is without doubt the model which has received to most publicity, out of the models that relate stock returns to explanatory variables. The theory was the first to explicitly define how much a company's "cost of equity" - read "return" - should exceed a benchmark rate. The theory shows that that under certain restricted set of conditions, a stock's expected excess return over a theoretical risk-free rate is a linear function of the expected market premium. The market premium is defined as the excess market return over the risk free rate. The CAPM equation is

$$(E(r_n) - r_f) = \beta_n E(Market - r_f)$$

where

E() is the expectation operator.

r_n is the return of stock n. Note that r_n is a random variable.

r_f is the return of a traded risk-free asset.

$Market$ is the market rate of return, which is defined as a weighted sum of all tradeable assets. Hence, $Market - r_f$ is the market risk premium.
β_n is defined to be $cov(r_n, Market)/var(Market)$.

The same expected return equation can be derived using the Arbitrage Pricing Theory (APT). For exposition purposes, the APT starts by considering a single-factor model. Uncertainty in the level of returns has two sources: a macroeconomic factor, which affects all firms, and a firm-specific effect. The macro, or common factor is assumed to have zero expected value, and is used to measure new information concerning the economy. Stock returns therefore satisfy the following probabilistic equation, relating return to the level of a macro-factor F, not necessarily the market return.

$$r_n = E(r_n) + \beta_n F + \epsilon_n$$

where
r_n is the return of stock n. Note that r_n is a random variable.
F is the macro-factor, has zero mean.
β_n, to be estimated by a time-series regression of return on the factor F.
ϵ is the error term, which is firm-specific.

In the case where the single macro-factor is the market return,the APT relies on the fact that if the preceding equation holds, and the CAPM equation does not hold for any well diversified portfolio, then an arbitrage - risk free profit - opportunity emerges, which instantly reestablishes the equilibrium in favor of the CAPM equation. A well-diversified portfolio means a portfolio where the idiosyncratic component of each stock's return has been made relatively small, by making the weight of any one stock in the porfolio low enough. The next step in the theory involves showing that if the CAPM equation is true for any well-diversified portfolio, then it must also be true for individual stocks.

The preceding discussion can be extended to more than one factor using the same arbitrage argument. For the two factor case, we assume the following return generating process holds.

$$r_n = E(r_n) + \beta_{n1}F_1 + \beta_{n2}F_2 + + \epsilon_n$$

where

r_n is the return of stock n. Note that r_n is a random variable.

F_1 and F_2 are macro-factors, with zero mean.

β_{n1} and β_{n2}, to be estimated by a time-series regression of return on the factors.

ϵ is the error term, which is firm-specific.

We then get the following expected return equation, for a portfolio with betas β_{P1} and β_{P2} .

$$(E(r_P) - r_f) = \beta_{P1}E(r_1 - r_f) + \beta_{P2}E(r_2 - r_f)$$

where

r_P is the return of portfolio P.

r_f is the return of a traded risk-free asset.

r_1 is the return on a portfolio with beta equal to one for the first factor, and zero for the second.

r_2 is the return on a portfolio with beta equal to one for the second factor, and zero for the first.

β_{P1} and β_{P2} the sensitivities of portfolio P with respect to factors F_1 and F_2 respectively.

It can also be shown[1] that if the preceeding relationship holds for all well-diversified portfolios, then it must hold for all stocks also.

Note that throughout our review of the CAPM and APT, no mention is made of correlation between stocks except that caused by the market return or any common factors.

3.1.2 Single Factor Models and Multifactor CAPM's

One of the drawbacks of refering to the CAPM equation when doing empirical work is its reliance on the unobservable market premium described above. A more general model replaces

[1] Ross (1976).

3.1.2 Single Factor Models and Multifactor CAPM's

One of the drawbacks of refering to the CAPM equation when doing empirical work is its reliance on the unobservable market premium described above. A more general model replaces the unobservable market premium with an observable market index. This market index is simply a weighted average return calculated over a sample of stocks present in the market. Usually, a capitalization weighted index such as the Standard and Poor's index is used. In the single factor model a stock is modeled as a linear function of the market index plus an error term as follows

$$r_n = \alpha_n + \beta_n R_m + \epsilon_n$$

where
r_n is the return of stock n.
R_m is the market index rate of return.
α_n and β_n are parameters to be estimated.
ϵ_n is an error term.

Though similar to the CAPM, the single factor model assumes that the market is a proxy for the combination of factors which in fact make returns fluctuate. These factors are macroeconomic in nature, and affect all firms. They might include business cycles, inflation, money-supply changes, technological changes, or prices of raw materials. The market index then serves as a macroeconomic indicator reflecting the levels of these factors.

The single factor model also assumes that the error terms are uncorrelated between stocks, so that the market index return R_m uniquely determines the correlation betweem stocks. Specifically, any two stocks a and b in the market will have a covariance equal to

$$Cov(r_a, r_b) = \beta_a \beta_b Var(R_m)$$

Hence, in the single factor model, conditional on R_m, returns can be modeled as uncorrelated.

Just as the APT allowed for more than one factor in the return generating process, other factors besides the market can be used in a regression of stock returns. The APT itself does not specify which factors should be included in such a regression, though further studies have dealt with the issue of choosing variables. Typically, macroeconomic variables are used. We then get the following type of regression equation, where the explanatory variables are observable macroeconomic indices, proxying for the underlying factors that would fit into an APT model. For example, unanticipated changes in the return on debt securities is a proxy for an interest rate factor, and the unanticipated changes in the value of the US dollar is a proxy for a factor capturing export sensitivity. We then get the following equation, where we assume that the error term is uncorrelated across time.

$$r_n(t) = E(R_n(t)) + \beta_{n1}F_1(t) + \beta_{n2}F_2(t) + \cdots + \epsilon_n(t)$$

where
(t) is the time index.
r_n is the return of stock n. Note that r_n is a random variable.
F_1 and F_2 are macro-factors, with zero mean.
β_{n1} and β_{n2}, to be estimated by a time-series regression of return on the factors.
ϵ is the error term, which is firm-specific.

Elton et al. [2] argue that the factors should in fact be expressed in terms of deviations from their expectation, with expectation meaning the forecasts given by professional analysts. Their rational is that only unexpected variations in these factors, or the proxies thereof, should affect stock returns.

3.1.3 Using Firm Specific Information

When the factors affecting returns are unobservable, with no available proxies, Rosenberg (1974) has developed an alternative approach which uses observable characteristics of the firm, also called "firm-specific" variables in this paper. His model starts with the hypothesis that returns can be described by a multifactor equation as follows.

[2] Elton, Edwin J., et al., "Cost of Capital Using Arbitrage Pricing Theory: A Case Study of Nine New York Utilities," Financial Markets, Institutions and Instruments, V.3, N.3.

$$r_n = \Sigma_{k=1}^{K} \lambda_{nk} F_k + \epsilon_n$$

where

r_n is the return of stock n.

F_k are unobservable factors, where K is the number of factors.

λ_{nk} is the sensitivity of stock n to factor k.

ϵ_n is the error term, which is firm-specific.

Notice that the factors are the same for all stocks in a given time period t. Also, the model works with panel data, in a one-period environment. The model for all stocks can be written in matrix form as

$$r = \Lambda F + \epsilon$$

where

r is the N vector of returns.

F is the K vector of unobservable factors.

Λ the $N \times K$ matrix of sensitivities.

ϵ is the N vector of sensitivities with variance σ_n^2.

The firm-specific variables enter the model in the following two equations. Specifically, the variance of the firm-specific error term and the sensitivities to the factors are assumed to be functions of characteristics of the firm. We have

$$\sigma_n^2 = \Sigma_j c_j x_{nj} + u_n = c' x_n + u_n$$

and

$$\lambda_n = D x_n + \epsilon_n$$

where

σ_n^2 is the variance of the return of security n. .

x_{nj} is the value of the firm specific variable j for firm n

c is a J vector of coefficients.

D is a $K \times J$ matrix of coefficients.

u_n and ϵ_n are error terms.

Some algebra yields the simple equation

$$r_n = x_n' F^* + v_n$$

where

F^* are the factors to be estimated.

v_n is an error term.

The error term v_n is a little bit less than well behaved, but consistent estimates of the F^* can be obtained using an appropriate weighted least-squares procedure.

3.1.4 Concluding Remarks on Probabilistic Factor Models of Stock Returns - the Nature of Relevant Explanatory Variables

Connor (1995), in a review of factor models of stock return, employs the following useful categorization of models. He distinguishes the models depending on which explanatory variables they use, yielding the categories " macroeconomic factor models," "fundamental factor models," and "statistical factor models." The models are not mutually exclusive, and can be combined - for example, combine fundamental factors with macroeconomic variables. One can even say that the models capture the same effect, assuming, for example, that fundamental factors are proxying for macroeconomic effects, or statistical factors - really indices constructed from stock groupings - are capturing these same macroeconomic factors. The list of factors Connor uses in the macroeconomic and fundamental categories is given in the following table.

When all five macroeconomic factors are used simultaneously, the R^2 of the OLS regression of stock returns on the factor is 0.109. When all the fundamental factors are used simulateously, the R^2 jumps to 0.426. With just the statistical factors, the R^2 is 0.390. The marginal explanatory power of the macroeconomic variables, when added to any of the other two complete sets

Macroeconomic Variables: inflation, term structure (of interest rates), industrial production, default premium of corporate bonds, unemployment
Fundamental - Firm Specific - Variables: industry dummy variables - these correspond to a given industry classification -, variability in markets, success, size, trade activity, growth, earnings to price, book to price, earnings variability, financial leverage, foreign investment labor intensity, dividend yield

Table 3.1: Variables that May Affect Stock Returns

of variables, is negligeable, implying that the macroeconomic variables do not explain anything that the other variables cannot capture. A regression of returns on the industry dummies alone causes the R^2 to be 0.163. When added to all the other explanatory variables, the industry dummies make the R^2 increase by 0.18 . They by far have the most power of all the variables, when considered as a group. It is too bad that the criterion for comparing models was R^2, since such a measure does not adjust for the loss in predictive power which can accompany the overfitted models. A more appropriate measure of fit would have been an adjusted R^2, or the AIC, both of which include a penalty for the number of coefficients estimated.

Notice that the macroeconomic variables described above do not include the estimated stock betas - from the single factor model described above. The market return does explain a significant amount of returns in a time-series regression. However, used in a cross-sectional context, this explanatory power vanishes. This effect, or lack thereof, is documented in Fama and French (1992). They run regressions of stock return on explanatory variables, and find the following firm-specific variables to be significant:

size (stock price times shares), leverage (value of debt), earnings to price, book to market equity (book value of firm's common stock to market value).

They find that used alone or in conjunction with other variables, the beta from the single market factor model, estimated exogenously from past data, carries little information about average returns in a cross-sectional regression of returns. In combination with other variables, size and book-to-market equity carry most of the information on average returns.

3.2 Clustering Algorithms to Group Stocks - The State of the Art

In this section we review various methods that explicitely try to come up with industry classifications, while pursuing some optimization criteria. Specifically, the two methods we focus on are factor analysis and cluster analysis. A heuristic which uses the covariance matrix of returns to cluster stocks together will also be discussed.

3.2.1 Factor Analysis

An example of factor analysis applied to the study of stock returns is given in King (1966). His stated goal is to study the mutivariate behavior of stock price changes over time. To this effect he analyses monthly data on 63 securities from the NYSE, from the period June 1927 to December 1960. He works with the first differences in the logarithm of price over a total of 403 months. The basic random variable he considers is therefore $y_{it} = \log price_{jt} - \log price_{j,(t-1)} = \log \frac{price_{jt}}{price_{j,(t-1)}}$. Note that this is just the log of return, if one abstracts from any potential dividends. In his work, King adjusts these variables for stock splits and dividends, making them genuine logarithms of return. The securities he chooses to study fall into six distinct SEC categories: tobacco products, petroleum products, metals, railroads, utilities and retail stores. He postulates that the correlation between stock returns can be explained in terms of a weighted sum of market, industry, and company effects. He then sets out to test the extent to which the industry-like clusterings within his sample correspond to the six SEC categories.

The basic factor analytic model applied to stock returns would postulate that returns at time t, r_t are a linear function of unobservable factors f_t and a random unobservable unique term. In equation form this gives

$$r_t = \alpha + \Lambda f_t + u_t$$

where

r_t is an N × 1 vector of observed logarithm of returns

α is an N × 1 vector of means

f_t is a K × 1 vector of unobservable random factors

Λ is an N $\times$ K fixed matrix of unobservable factor coefficient loadings

u_t is an N $\times$ 1 vector of random unobservable unique terms

In addition, the standard factor analysis requires the following assumptions: $\mathrm{E}(u_t)=0$, $\mathrm{E}(u_t' u_t)=\Phi$, a diagonal matrix, and the elements of f_t are uncorrelated with those of u_t. Then, the covariance matrix of r_t can be written as $\Sigma = \Lambda \mathrm{E}(f_t f_t') + \mathrm{E}(u_t u_t') = \Psi + \Phi$.

King performs a factor analysis of the correlation matrix of returns, using a principal components type of estimation procedure. Note that the basic relationships explained above for the covariance matrix parallel those for the correlation matrix. He estimates the loadings, i.e. the columns of Λ, for seven factors, his a priori assumptions being that the first one represents a market factor, and the six other ones will represent industry factors. After rotation of the factor, he is able to find remarkeable agreement between the SEC categories, and the groups of stocks suggested by the analysis. As far as the goodness of fit of the model is concerned, he reports that the total communality explained by the factors - ratio of total variance explained by the seven factors to the sum of the variance terms of the returns - is equal to 0.863, which represents a relatively good fit. The most interesting fact as relates to this research is the agreement between the SEC categories, and the categories suggested by the factor analysis.

Another study which uses factor analysis to explain the covariance matrix of stock returns was conducted by Lehman and Modest (1993). Their motivation is different from King's, in that they are trying to test the APT, i.e. explain returns in terms of factors suggested by the factor analysis. To this end, they take the factors suggested by the factor analysis, and use them to construct factor portfolios, which are weighted combinations of stock returns. These factor portfolios can in turn be used as explanatory variables of return, in a regression setting.

They test the APT by examining whether the theory can explain "well documented anomalies": the fact that returns seem to be dependent on variables such as firm size and dividend yield. This fact is called an anomaly because it lacks economic interpretation. Their data consists in 750 stocks, tracked weekly over four periods:1963-1967, 1968-1971, 1973-1977, 1978-1982. In all their are 403 weekly observations of the 750 stock returns. They estimate the factor loadings using a maximum likelihood procedure. They successively consider models with 5, 10, and 15 factors. Their tests of the APT involve constructing portfolios ranked on firm size or dividend yield, and then estimating the parameters of a regression of these ranked portfolio returns on the space of factors suggested by the factor analysis. They then test whether the intercept terms in these regressions are significantly different between portfolios ranked at different ends of the spectrum - either in terms of size or dividend yield. They conclude that the intercept terms are different between such extreme portfolios, and so reject the hypothesis that their factors explain returns completely. This study relates to our present research because the factors are constructed to be perpendicular to each other, and could be interpreted as indexes for different industry sectors. Given the size of our dataset, however, and because we are trying to use the a priori information contained in the industry classifications contained in our datasets, factor analysis did not seem like a practical alternative to implement on our PC, given memory restrictions in the software SPLUS[3] that we used.

3.2.2 Pure Clustering Algorithms

The literature is very thin on this topic. Clustering stocks does not seem to have received much coverage. The following two papers are noteworthy applications of clustering techniques to financial data. The first deals with the grouping of stocks on the basis of correlation with other stocks, and the second deals with the grouping of mutual fund according to management style.

Farrell (1974) uses a method inspired by King (1966), called step-wise clustering. From a statistical standpoint, the method is, like most clustering algorithms, not rigorous in that it does not lend itself to rigorous testing. King dubs it "quick and dirty factor analysis." Farrell works with a sample of 100 major stocks, all of which are listed on a national exchange, and 90 of which belong to the SP 500. The stocks also span 60 of the SP 500 industry classes. The returns are monthly returns covering the years 1961-69.

He begins by regressing the returns from his 100 stock sample onto the SP 425 stock market index monthly rate of return. He then works with the covariance matrix of residuals from the single market index model, and iteratively groups stocks two at a time using a "highest correlation" criterion. Specifically, the algorithm he uses involves three steps: (1) searching the residual covariance matrix for the two variables with the highest positive correlation coefficient. (2) combining these variables to reduce the dimension of the matrix by one. (3) recomputing the

[3] Splus, Version 3.1.

correlation matrix to include the correlation between the combined variable and the remaining variables.

Even though the stepwise clustering method is just a heuristic implementation of his stated goal, to obtain groups of returns such that intra-group residual correlation is low, but inter-group residual correlation is high, he validates his results in several ways. First, he obtains groupings which correspond to his prior assumptions, namely that the stocks can be categorized into (a) growth stocks (b) cyclical stocks (c) stable stocks. In addition, he finds a fourth group consisting entirely of stocks from the petroleum industry. Hence the addition to his list of the category (d) oil stocks. He examines then orders the stock according to their group identity, and examines the submatrices of within group covariance and in between group covariance. He notices a predominance of positive correlation coefficients, as well a a greater number of significantly positive correlation coefficients than would be expected by chance. Hence stocks within groups are highly correlated. He then goes on to notice the large number of negative correlation coefficients in the in-between covariance matrices, and lack of significantly positive correlation coefficients. This indicates a low degree of correlation across groups.

He uses another validation procedure which involves testing for the significance of correlation between the residuals of different regressions. The first regression is that of return on the market index. The four others involve regressing the group average return on the market index return. The hypotheses he tests and fails to reject are (a) that the residuals of the first regression and any of the other four are highly correlated- note that this involves looking at 4×100 correlations - and (b) that the correlation between the residuals of the last four regressions are uncorrelated - this involves looking at six correlations.

Brown and Goetzman (1995) use a classification algorithm to group mutual funds into groups of similar style. Style refers to the management style, with typical examples being "growth" and "value". They motivate their study by noting that institutional reported styles are typically misleading. Managers can self report which style category they are in, and can change categories if it will make their performance look better. There is therefore a need for a stylistic classification that is empirically and objectively determined. The movement of mutual fund returns seems to be explained much better by the empirically determined style indexes which Brown and Goetzman determine.

Their study yields style categories which are not consistent with institutional classifications, but do a better job of predicting cross-sectional variation in fund return. They use a cross-sectional time-series of mutual fund returns from 1976 through 1994. They start by postulating the existence of K styles, and write returns as $r_{nt} = \alpha_{jt} + \beta_{jt} + \epsilon_{jt}$, where fund n belongs to style j, and I_t are explanatory variables, which could be macroeconomic variables, or the value of indices. The algorithm assigns funds to styles on a per-period basis, by minimizing the sum squared of errors in the above model.

Chapter 4

Data

In this section we describe the data on which we test our probabilistic models and clustering algorithms.

4.1 Description of the datasets.

The data consists in monthly datasets for the Japanese stock market, spanning the period February 1988 to August 1995 - almost eight years. We also had available datasets for each of the G7 countries except the USA; that is, for Germany, France, Italy, the UK and Ireland, Japan, and Canada. But we restrict our research to the Japanese datasets. A monthly dataset contains the following list of 23 variables for an average of 2000 companies : Ticker, Company name, Country code (in this case, the code is for Japan), Currency in which company stock is traded (we restrict ourselves to stocks traded in Yen) , Price in the previous currency , Monthly return (a percentage), Currency return relative to US dollar, Vestek code, Worldscope code, IBES industry code, Local market code (only for France, Italy, UK and Japan), PE ratio (price to earnings, i.e. price per share divided by earning per share), Capitalization (total market equity of stock; this number is given in millions in the original dataset), Yield, PB ratio (price to book, i.e.price per share divided by book value per share), Earnings per share, Total dividend in month, Volume traded (during that month), Shares outstanding (number of shares of stock), Book value (in millions, in the original dataset), Total liabilities, Total assets, Sales (in the currency in which the stock is traded). The variable volume, however, is zero or not available in all the datasets, eliminating it as a useful tool for future analysis.

4.2 Choosing Variables to Include in the Analysis

We choose to work with four variables out of the original 23. These are: Return, Capitalization, PE ratio, Book value. The variable Capitalization was given in millions of the currency unit in which the stock is traded, so that we multiply the variable by 10^6before we actually use it to create the following other variables. Note that the variable Capitalization could have been reconstructed by using the formula Capitalization = Shares × Price. Also, the variable PE ratio could have been created by using the formula PE ratio = Price / Earnings per share. Out of the original four variables, as stated in the next section, where we talk about screening the data to prepare it for our work, we construct three variables.

ME = log(Capitalization)

E/P = 1/(PE Ratio)

BE/ME = log(Book value/Capitalization)

We use these last three variables in the following analysis of the data, and in our subsequent model building, and clustering algorithm implementations.

4.3 Screening of the datasets.

The data - only the four variables which we decided to include in our analysis - were analyzed for aberrant observations such as negative valued observations where only positive values have any logical significance. A number of functions written in SPLUS further "screen" the data to make it conform to our working requirements. The following list describes the purpose and extent of the preliminary analysis of the datasets.

1. Eliminating foreign currencies.

   For Japan, companies not traded in Yen are discarded.

2. Eliminating negative valued observations.

The following variables are examined for negative values, and in the event where a negative value is found, are to be treated as a missing value:

- Capitalization
- PE Ratio
- Book Value

1. Eliminating false zeros.

   For some variables, zero values cannot be interpreted, and so they will be treated as missing values. These variables are:

   - Capitalization
   - PE Ratio
   - Book Value

2. Eliminating "penny stocks".

   Here, we remove stocks with the 5% lowest capitalization. Our justification is that these stocks tend to not be indicative of the general market behavior, with their distribution of returns being too noisy.

3. Transforming the data to obtain three relevant variables.

   As mentionned previously, we take as given that three variables, market equity (ME), earnings to price (E/P), and book-to-market ratio (BE/ME) are determinants of price. We only consider these variables in our empirical work. We construct them from our data according to the following rules:

   ME = log(Capitalization)
   E/P = 1/(PE Ratio)
   BE/ME= log(Book Value/Capitalization)

A *sector variable* is defined as the average of the variable across the stocks within a sector. The average of a *sector variable* across all time periods when it was available is the *average sector variable*. In the next 6 tables, we show the *average sector variable* for the 69 Vestek industry sectors, when applicable - that is, if there were more than two observations, for more than two time periods, so that the term "average" has a meaning, both to calculate the sector variable,

Vestek Industry Code	Return	SDV	ME	E/P	BE/ME	Number of Observations
1	-0.3337	6.362	24.24	0.09038	-0.5533	70
2	-0.328	6.228	24.97	0.03174	-1.499	71
3	-1.292	7.606	23.01	0.1584	0.2236	65
4	-0.07396	5.952	26.11	0.1815	-0.3814	77
5	-0.2309	7.283	24.6	0.1115	-0.2909	71
6	NA	NA	NA	NA	NA	NA
7	0.1406	6.875	24.76	0.6451	0.04414	71
8	.0.3721	7.924	24.67	0.1014	-0.65	77
9	0.3511	7.748	24.83	0.08538	-0.7401	77
10	0.2604	7.634	24.72	0.1076	-0.4635	77

Table 4.1: Vestek Industry Sectors 1 through 10 and Their Corresponding Variables

Vestek Industry Code	Return	SDV	ME	E/P	BE/ME	Number of Observations
11	0.4566	7.877	24.72	0.08029	-1.008	71
12	0.7815	8.792	24.68	0.1217	-0.5737	77
13	NA	NA	NA	NA	NA	NA
14	-0.01597	7.55	24.22	0.08289	-0.2543	71
15	-0.3819	7.517	24.44	0.1613	0.08212	59
16	0.694	8.238	24.59	0.1419	-0.4208	77
17	0.3557	9.492	24.31	0.1521	-0.4926	77
18	NA	NA	NA	NA	NA	NA
19	0.04558	7.789	23.59	0.1943	-0.07122	57
20	0.5975	8.103	24.59	0.1432	-0.4759	77

Table 4.2: Vestek Industry Sectors 11 through 20 and Their Corresponding Variables

and to calculate its average. If the calculation of the sector variables was not applicable to any sector, its row is filled with the sign NA, which stands for not available. The time periods that we consider here are February 1988 to August 1995, with the exception of April 1989. January 1992 , and May 1995 - we did not include these dates because the corresponding datasets had a large number of errors. In the following pages, we present some relevant graphs pertaining to the distribution the variables within sectors. Precisely, we provide the histograms of average sector return. average sector return standard deviation (SDV), average sector ME. average sector E/P, and average sector BE/ME.. We then provide plots of sector return with each one of the other variables. No clear relationship is apparent between sector return and the sector variables.

Vestek Industry Code	Return	SDV	ME	E/P	BE/ME	Number of Observations
21	0.5209	9.145	23.84	0.1091	-0.8726	71
22	NA	NA	NA	NA	NA	NA
23	NA	NA	NA	NA	NA	NA
24	-0.1638	4.697	26.37	0.09851	-0.8807	71
25	0.2159	6.923	24.14	0.07722	-0.4151	77
26	0.311	8.781	24.36	0.125	-0.6311	77
27	NA	NA	NA	NA	NA	NA
28	-0.9796	6.936	23.37	0.1056	-0.6877	59
29	0.616	8.505	24.45	0.1073	-0.5078	77
30	1.279	9.076	23.16	0.03926	-0.3846	47

Table 4.3: Vestek Industry Sectors 21 through 30 and Their Corresponding Variables

Vestek Industry Code	Return	SDV	ME	E/P	BE/ME	Number of Observations
31	0.2372	7.623	24.9	0.1136	-0.4167	77
32	NA	NA	NA	NA	NA	NA
33	0.3501	8.3	24.88	0.07959	-0.9058	77
34	0.8481	7.934	24.72	0.172	-0.6012	77
35	NA	NA	NA	NA	NA	NA
36	-0.02967	7.756	24.68	0.06901	-0.7587	77
37	0.3369	6.85	25.52	0.1353	-0.8248 7	7
38	NA	NA	NA	NA	NA	NA
39	0.03666	6.875	25.23	0.1275	-0.461	5 77
40	NA	NA	NA	NA	NA	NA

Table 4.4: Vestek Industry Sectors 31 through 40 and Their Corresponding Variables

Vestek Industry Code	Return	SDV	ME	E/P	BE/ME	Number of Observations
41	0.6734	8.551	24.1	0.1613	-0.5304	77
42	0.1893	8.988	24.3	0.4053	-0.9856	77
43	NA	NA	NA	NA	NA	NA
44	-0.4281	7.271	25.24	0.2594	-0.1338	77
45	-0.008967	7.465	24.52	0.09681	-0.5866	77
46	-0.0861	7.212	24.8	0.143	-0.5962	5
47	0.6221	4.94	25.46	0.1732	-0.3331	71
48	0.4164	8.141	24.19	0.1197	-0.4624	77
49	NA	NA	NA	NA	NA	NA
50	0.2589	7.939	23.69	0.1209	-0.495	71

Table 4.5: Vestek Industry Sectors 41 through 50 and Their Corresponding Variables

Vestek Industry Code	Return	SDV	ME	E/P	BE/ME	Number of Observations
51	0.1074	8.128	24.71	0.208	-0.4646	77
52	-0.2142	7.105	25.04	0.1458	-0.1121	77
53	0.5092	8.278	25.03	0.1957	-0.5325	77
54	0.1159	8.123	24.92	0.144	-0.5459	77
55	0.7244	6.162	25.4	0.09608	-0.7952	70
56	-1.566	7.668	24.75	0.06908	-0.2688	40
57	0.2093	7.889	24.37	0.09539	-0.8148	77
58	0.1747	5.689	26.11	0.03557	-1.468	77
59	-0.06067	7.431	25.4	0.1303	-0.705	77
60	NA	NA	NA	NA	NA	NA

Table 4.6: Vestek Industry Sectors 51 through 60 and Their Corresponding Variables

Vestek Industry Code	Return	SDV	ME	E/P	BE/ME	Number of Observations
61	0.2369	9.849	24.17	0.09436	-0.8725	77
62	NA	NA	NA	NA	NA	NA
63	0.1189	8.223	24.35	0.1539	-0.3329	77
64	-0.03053	6.529	25.49	0.07363	-1.117	77
65	0.3275	3.794	27.38	0.08807	-1.164	56
66	-0.04067	8.142	24.58	0.0987	-1.727	77
67	-0.2385	2.392	27.44	0.08644	-0.4046	77
68	-0.2357	5.918	25.77	0.03004	-1.025	71
69	-1.181	3.701	26.91	4.161	0.8495	50

Table 4.7: Vestek Industry Sectors 61 through 69 and Their Corresponding Variables

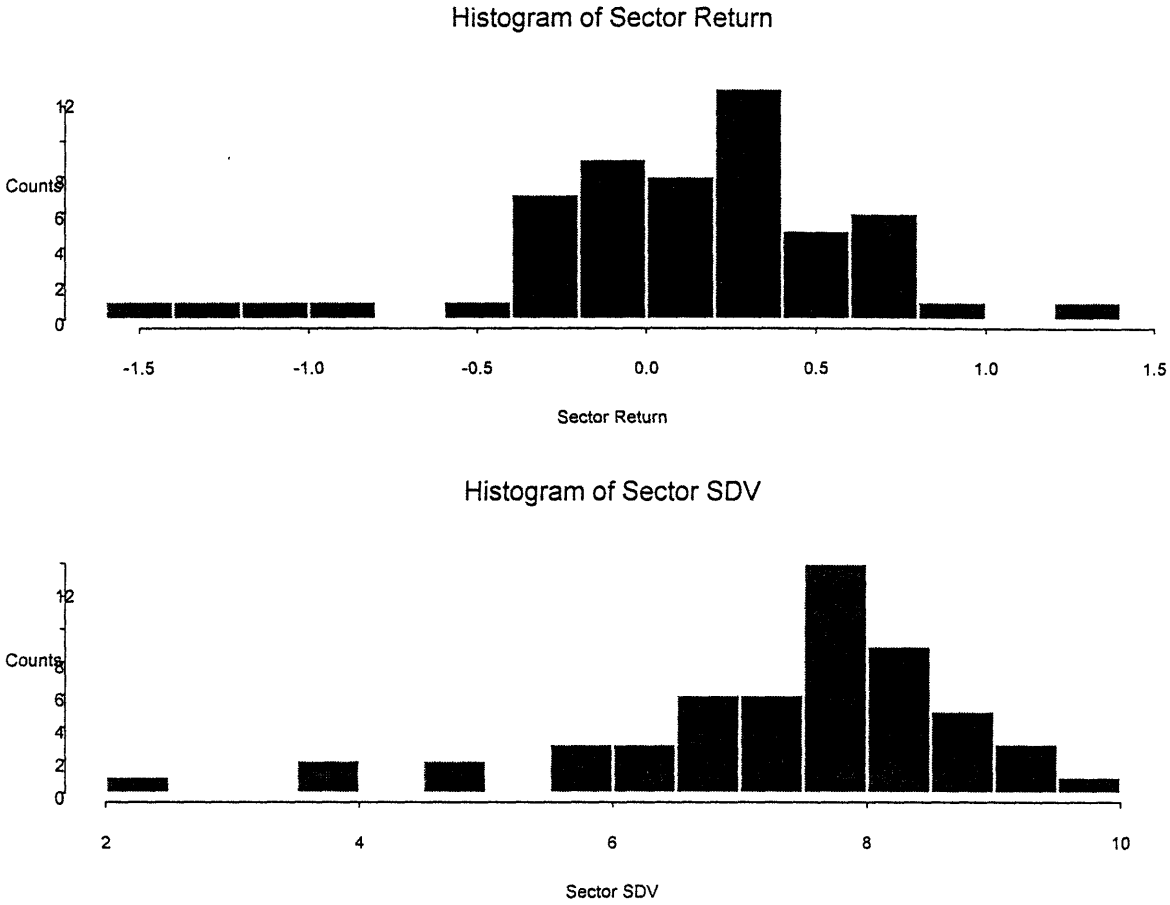

Figure 4-1: Histograms of Sector Return and Sector SDV

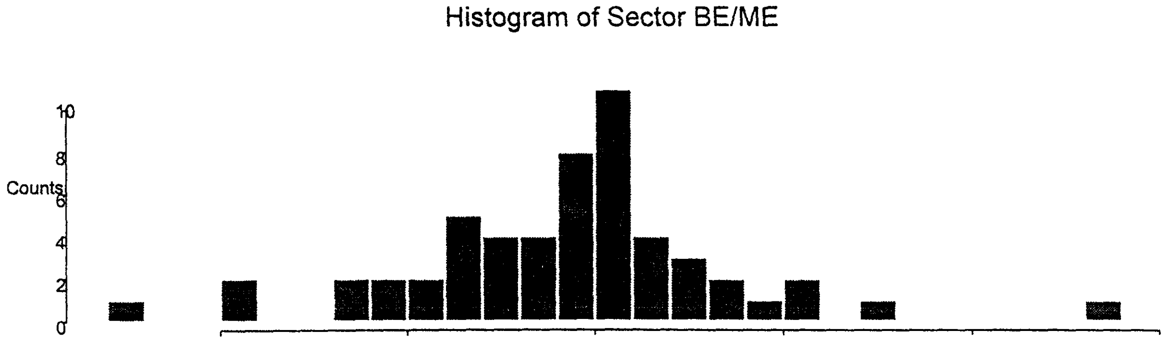

Figure 4-2: Histogram of Sector BE/ME

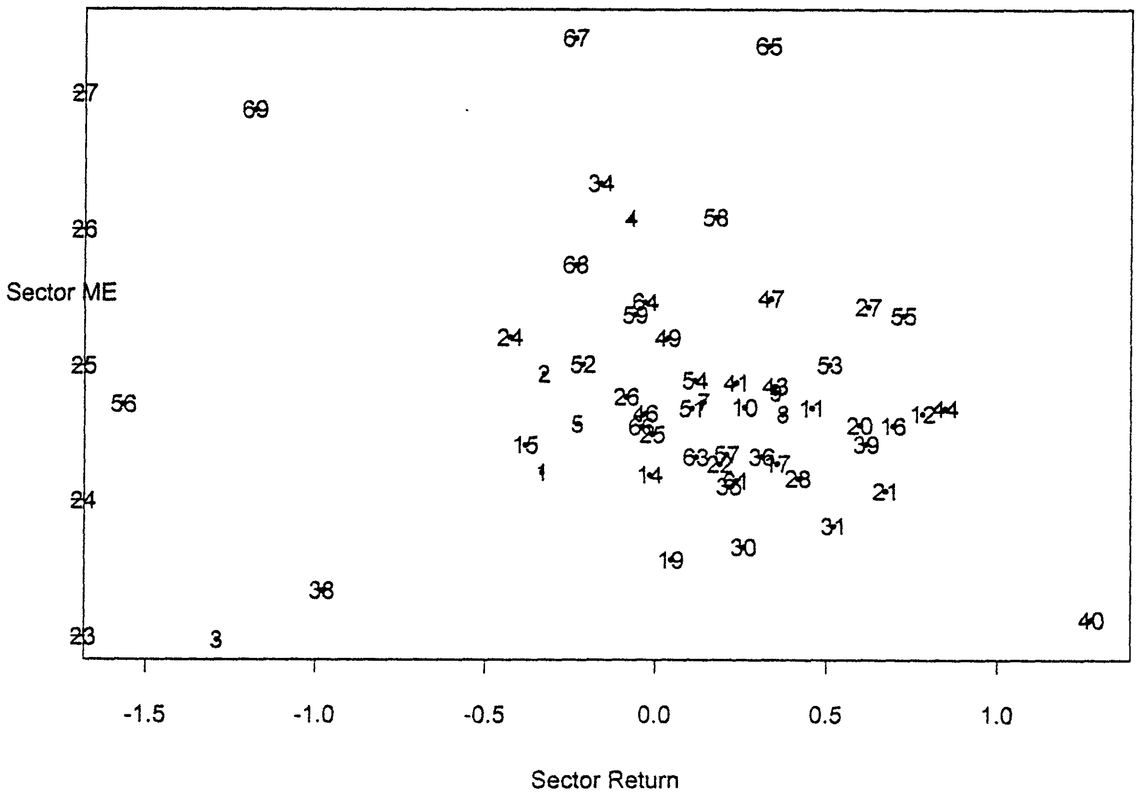

Figure 4-3: Plot of Sector Return and Sector ME (the numbers next to the points are Vestek Industry Codes)

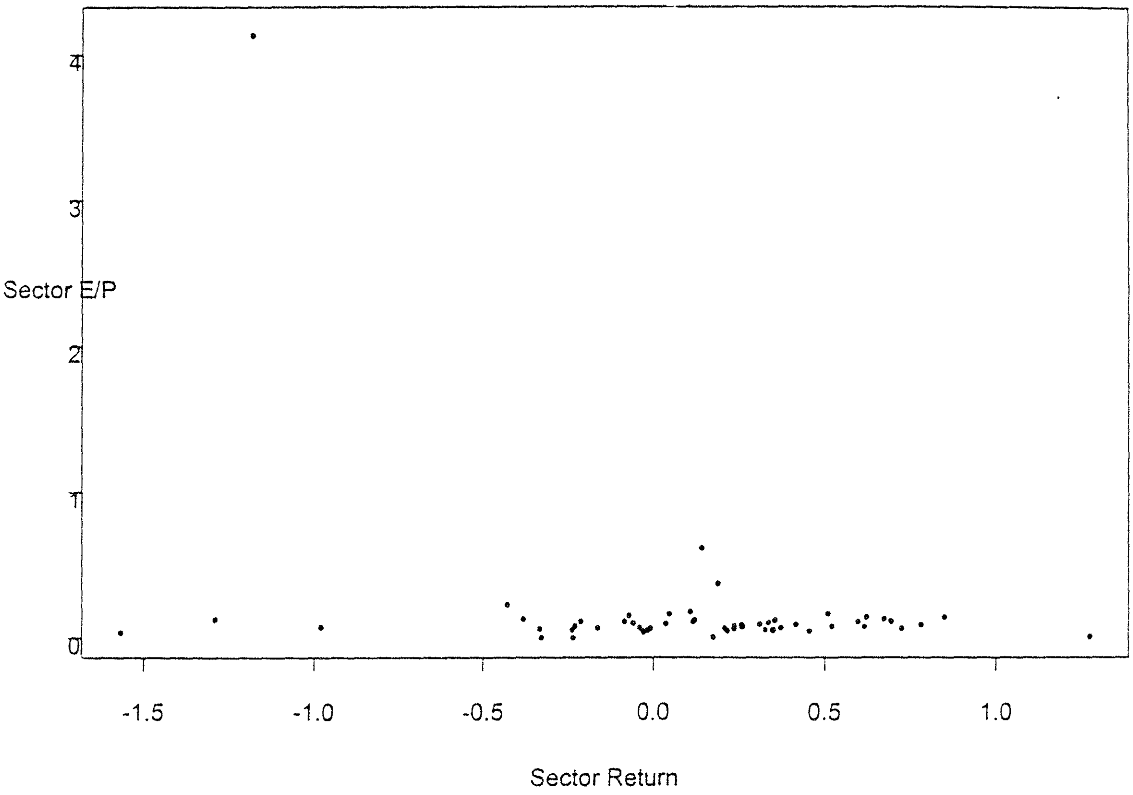

Figure 4-4: Plot of Sector Return and Sector E/P

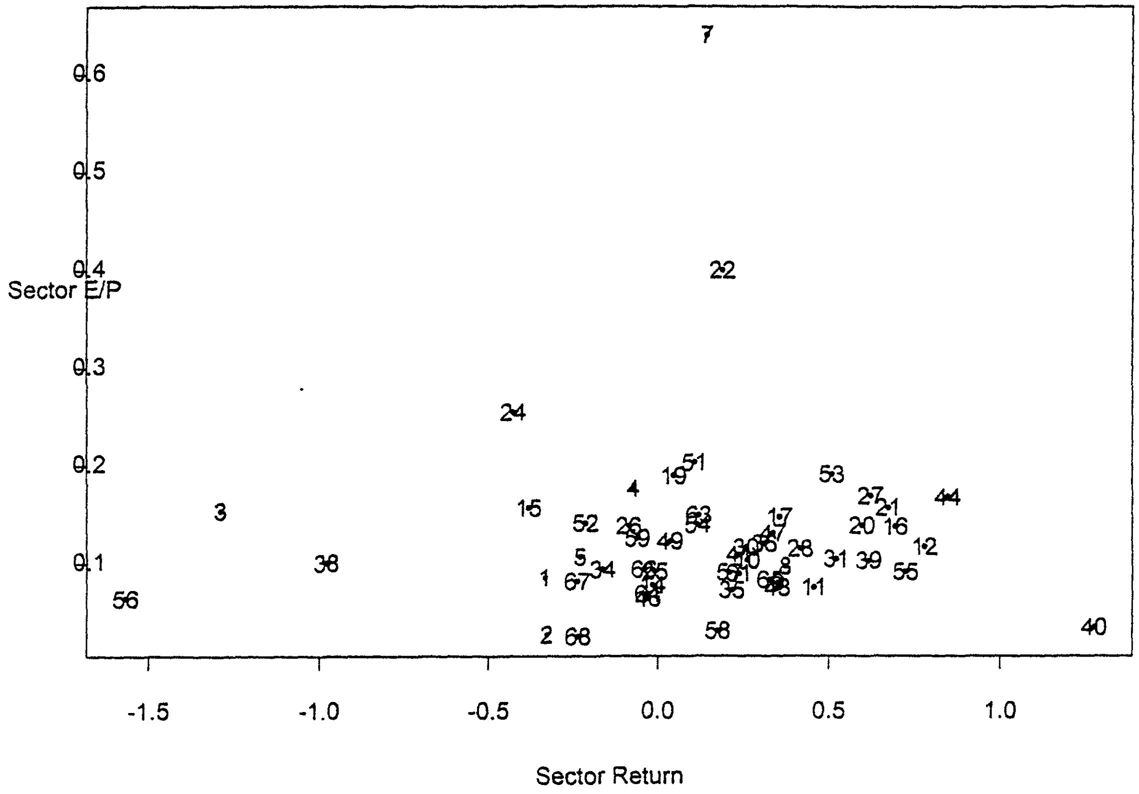

Figure 4-5: Plot of Sector Return and Sector E/P (the numbers next to the points represent Vestek Industry Codes). Note: Sector 69 was removed for scaling purposes.

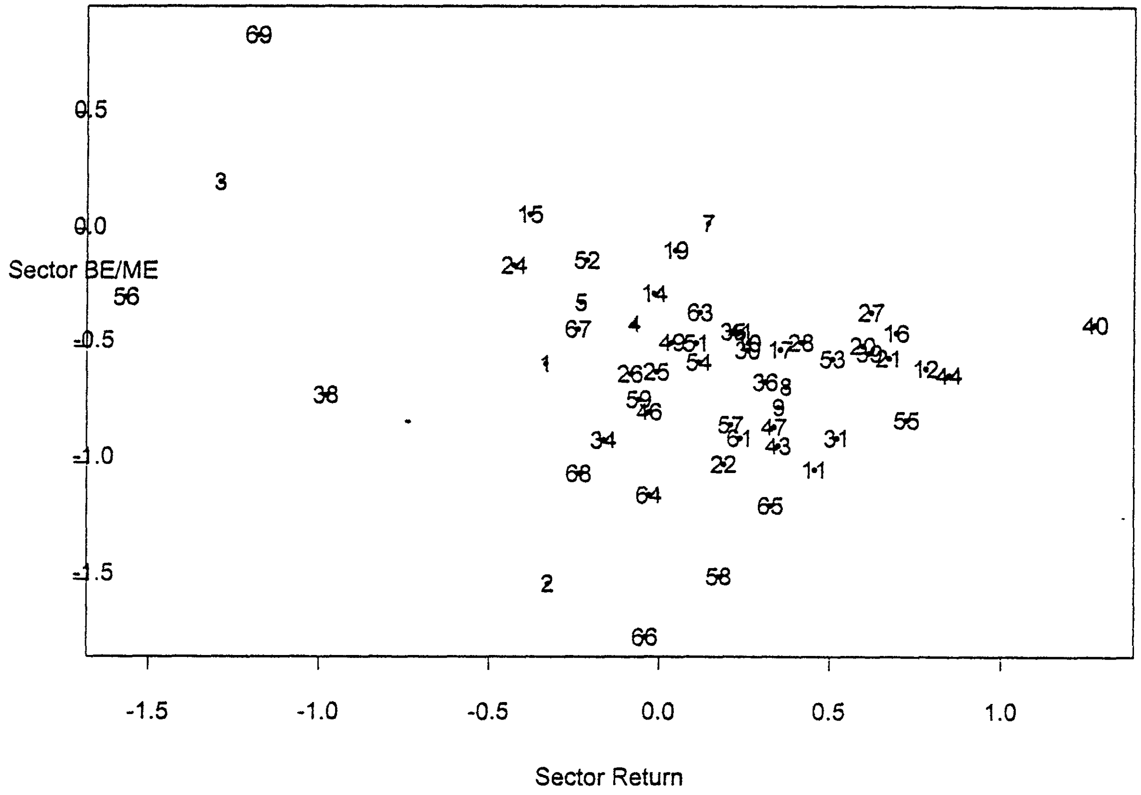

Figure 4-6: Plot of Sector Return and Sector BE/ME (the numbers next to the points represent Vestek Industry Codes)

Chapter 5

Probabilistic Framework

In this section we supply the details of our parametric probabilistic models of stock returns. The models relate stock returns and explanatory variables for a given month. We start by describing the most general form of the models we consider, where a stock's mean return is a linear function of firm-specific explanatory variables, and where its variance is the exponential of a linear function of the explanatory variables. If the linear function of the variance is a constant across all stocks and sectors, we have homoskedasticity and an OLS regression framework. If we constrain the linear function in the the variance term to have parameters equal to zero, but allow the variance to vary between sectors, we satisfy the assumptions of the groupwise heteroskedastic model.

5.1 Model in Its Most General Form

Given the firm-specific variables and the industry classification, we model stock returns normally distributed as in the following equations.

$$r|X \sim N\left(\mu(X,C), \Sigma(X,C)\right), \tag{5.1}$$

where

r is the N vector of stock returns.

X is the $N \times K$ matrix of firm specific explanatory variables.

C is the $N \times J$ industry classification matrix. Each column j corresponds to a sector, and any row n has a unique 1 in the column corresponding to the sector where the firm n belongs.

$\mu(X, C)$ is the N mean vector.

$\Sigma(X, C)$ is the $N \times N$ covariance matrix.

N is the number of stocks in time period t

K is the number of firm-specific explanatory variables we are considering

J is the number of sectors in the given industry classication C.

Note that we are allowing both the mean vector μ and the covariance matrix Σ to depend on the explanatory variables X and the industry classification C.

5.1.1 Assumption of Conditional Independence Across Stocks

We arrange the stock returns in the vector r so that the returns of all stocks in a same sector are adjacent. Specifically, we can write, denoting by J the number of sectors in our model,

$$r = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_J \end{bmatrix}. \tag{5.2}$$

where

r_j is the N_j vector of returns of group j, where $0 \leq j \leq J$,

N_j is the number of stocks in sector j.

Rearranging the matrix X accordingly we get

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_J \end{bmatrix} . \qquad (5.3)$$

where

X_j is the $N_j \times K$ vector of returns of group j, where $0 \leq j \leq J$.

We assume from now on that returns during a given period are independent conditional on the firm-specific explanatory variables and on the industry classification. This is an assumption which is standard in the literature concerning the description of cross-sections of stock returns[1]. Note, however, that we allow the industry classification to affect both the mean and variance of returns, as shown above. We are therefore capturing similarities in the probabilistic behavior of returns which are in the same sector.

With returns independent and therefore uncorrelated between stocks, we model the covariance matrix $\Sigma(X, C)$ of the population return vector r as diagonal as follows.

$$\Sigma = \begin{bmatrix} \Sigma_1 & 0 & \dots & 0 \\ 0 & \Sigma_2 & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \dots & \Sigma_J \end{bmatrix} , \qquad (5.4)$$

where Σ_j is the diagonal covariance matrix of sector j.

Each individual group j defines a model described by

$$r_j | X_j \sim N\left(\mu_j(X_j), \Sigma_j(X_j)\right), \qquad (5.5)$$

where

[1] see, for example, Fama and French, 1992

j integer, $0 \leq j \leq J$, where J is the number of groups in the model.

r_j is the N_j vector of stock returns.

X_j is the $N_j \times K$ matrix of explanatory variables for group j.

$\mu_j(X_j)$ is the N_j mean vector for sector j.

$\Sigma_j(X_j)$ is the diagonal $N_j \times N_j$ covariance matrix for sector j.

Notice that each group has the same number K of explanatory variables.

Given the form of the covariance matrix Σ above, the density function for the model is the product of the densities for the J groups defined by the sectorization. We can write

$$p(r|X) = \prod_{j=1}^{J} p(r_j|X_j)$$

where

r is the N vector of stock returns.

X is the $N \times K$ matrix of explanatory variables.

r_j is the N_j vector of stock returns.

X_j is the $N_j \times K$ matrix of explanatory variables for group j.

5.2 Regression Framework: Specifying $\mu_j(X_j)$ and $\Sigma_j(X_j)$

We here specify in its most general form the structure of the underlying mean vector μ_j and covariance matrix Σ_j of group j. We later test whether restrictions on this general form lead to a better model specification. The regression framework is described below. We allow the parameters of our model to vary between groups.

The return vector for group j is expressed as:

$$\begin{aligned} r_j &= \mu_j + \epsilon_j & (5.6) \\ &= 1\alpha_j + X_j\beta_j + \epsilon_j & (5.7) \end{aligned}$$

where

X_j is an $N_j \times K$ matrix of explanatory variables.

β_j is a K vector of parameters, specific to group j.

α_j intercept term for group j.

$\mathbf{1}$ is an N_j vector of ones.

ϵ_j is an N_j vector of errors, assumed to be jointly normally distributed $N(0, \Sigma_j(X_j))$.

We further specify the covariance matrix structure by defining the elements diagonal matrix $\Sigma_j(X_j) = diag(\sigma_j(n))$ of group j to be such that:

$$log(\sigma_j(n)) = \gamma_j + x_j(n)\theta_j \tag{5.8}$$

where

$x_j(n)$ is the row n of the $N_j \times K$ matrix of explanatory variables X_j of group j

γ_j is a constant term specific to sector j.

θ_j is a K vector of parameters, specific to group j.

This last relationship implies that

$$\sigma_j(n) = exp(\gamma_j + x_j(n)\theta_j) \tag{5.9}$$

which keeps the $\sigma_j(n)$ terms from being negative, which is reasonable since the terms $\sigma_j(n)$ represent variance terms. Notice again that the vector of paramters θ is constrained to be identical across sectors. However, we allow the variance to change between sectors and observations. This fact makes our model more general than other models of stock returns, which assume constant variance across observations.

5.3 Loglikelihood and AIC Calculation

Because return between sectors are independent, the loglikelihood function of the model is the sum of the loglikelihoods of the models for each sector. Of course, since the independence really is between observations, we can decompose the loglikelihood further, as shown below.

$$\begin{aligned}
& \ell(\mu(X,C),\Sigma(X,C)) \qquad (5.10)\\
= & -\frac{1}{2}\left(N\log 2\pi + \log|\Sigma| + (r-\mu)^T(\Sigma)^{-1}(r-\mu)\right)\\
= & -\frac{1}{2}\sum_{i=1}^{J}\left(N_j\log 2\pi + \log|\Sigma_j| + \left(r_j-\mu_j\right)^T(\Sigma_j)^{-1}\left(r_j-\mu_j\right)\right)\\
= & -\frac{1}{2}\left(\sum_{j=1}^{J}\left(N_j\log 2\pi + \sum_{n=1}^{N_j}\left(\log\left(\sigma_j(n)\right) + \left(r_j^n-\mu_j(n)\right)^T(\sigma_j(n))^{-1}\left(r_j^n-\mu_j(n)\right)\right)\right)\right)
\end{aligned}$$

where

P are the degrees of freedom of the parameters in the regression framework.

l is the loglikelihood of the model.

μ_j is the mean return vector of group j, given by the above equation relating returns and explanatory variables for each group.

Σ_j is the covariance matrix of the group j error vector ϵ_j .

$\mu_j(n)$ is the mean of the n^{th} stock in group j.

$\sigma_j(n)$ is the variance of the n^{th} stock in group j.

The parameters of the model are estimated by maximum-likelihood, where we maximize the above log-likelihood expression.

5.4 Mathematical Statement of the Problem

Once we have specified the form of the model to describe the probabilistic behavior of returns, we are left with the problem of choosing the best mergings of the original industry sectors. Specifically, we want to solve the following mixed integer program.

Minimize $\boldsymbol{AIC}$

$$= \sum_{j=1}^{J} \left(\log 2\pi + \sum_{n=1}^{N_j} \left(\log\left(\sigma_j(n)\right) + \left(r_n - \mu_j(n)\right)^T \left(\sigma_j(n)\right)^{-1} \left(r_n - \mu_j(n)\right) \right) \right) + 2 \times P$$

subject to

$$\mathbf{1(i,j)}(\alpha_i - \alpha_j) = 0$$

$$\mathbf{1(i,j)}(\gamma_i - \gamma_j) = 0$$

$$\mathbf{1(i,j)}(\beta_i - \beta_j) = 0$$

$$\mathbf{1(i,j)}(\theta_i - \theta_j) = 0$$

where

$\mu_j(n) = \alpha_j + x_j(n)\beta_j$ is the mean of stock n in sector j

$\sigma_j(n) = exp(\gamma_j + x_j(n)\theta_j)$ is the variance of stock n in sector j

$\mathbf{1(i,j)}$ is 1 if sectors i and j have been merged, 0 otherwise, with $0 \leq i \leq j \leq J$.

J is the number of original sectors

P is the number of independent parameters to be estimated.

Minimization of course takes place over the indicator variables $\mathbf{1(i,j)}$ and over the parameters to be estimated, the scalars α_j, γ_j, and the vectors β_j and θ_j,$1 \leq j \leq J$.

In the case where we want to solve a more general problem - which we do not consider here - where we allow for sector splittings as well as mergings, while allowing a maximum of J original sectors to exist, we have the following integer program.

Minimize $\boldsymbol{AIC}$

$$= \sum_{n=1}^{N} \left(\log 2\pi + \sum_{j=1}^{J} \mathbf{1}_{\mathbf{n}}^{\mathbf{j}} \left(\log\left(\sigma_j(n)\right) + \left(r_n - \mu_j(n)\right)^T \left(\sigma_j(n)\right)^{-1} \left(r_n - \mu_j(n)\right) \right) \right) + 2 \times P \quad (5.11)$$

subject to

$$\sum_{j=1}^{J} \mathbf{1}_{\mathbf{j}}^{\mathbf{n}} = 1$$

$$\mathbf{1(i,j)}(\alpha_i - \alpha_j) = 0$$

$$\mathbf{1(i,j)}(\gamma_i - \gamma_j) = 0$$

$$\mathbf{1(i,j)}(\beta_i - \beta_j) = 0$$

$$\mathbf{1(i,j)}(\theta_i - \theta_j) = 0$$

where

$\mu_j(n) = \alpha_j + x_j(n)\beta_j$ is the mean of stock n in sector j.

$\sigma_j(n) = exp(\gamma_j + x_j(n)\theta_j)$ is the variance of stock n in sector j.

$\mathbf{1}_{\mathbf{j}}^{\mathbf{n}}$ is 1 if stock n belongs to group j, 0 otherwise.

$\mathbf{1(i,j)}$ is 1 if sectors k and l have been merged, 0 otherwise, with $0 \leq i \leq j \leq J$.

J is the number of original sectors.

P is the number of independent parameters in the model.

Minimization takes place over the indicator variables $\mathbf{1}_{\mathbf{j}}^{\mathbf{n}}$,$\mathbf{1(i,j)}$ and over the parameters to be estimated,the scalars α_j, γ_j, and the vectors β_j and θ_j,$1 \leq j \leq J$. We will not attempt to find the optimal solution to any of the minimization problems stated above. They are there just to clarify the nature of the problem in its purest mathematical form. In other words, once we have specified the form of the model, we could theoretically find the best industry classification, that would minimize the AIC criterion.

5.5 Constraints on the Parameters of the Model: Special cases

In the next chapter, we implement the following four special cases of the more general formulation described above. We do not estimate the most general form of the model, which has too many parameters for practical estimation. Rather, we consider constraints on the original specification. Note that these models refer to the model that is estimated at each time period. The parameters from one period to the next are therefore assumed to be independent.

5.5.1 Homoskedasticity:OLS on Firm-Specific Variables and a Constant

If we constrain $\gamma_i = \gamma_j$ and $\alpha_i = \alpha_j$, $\beta_i = \beta_j$ across all sectors i,j, and $\theta_i = 0$ for all sectors i, estimating our model by maximum likelihood will give us the same parameter estimates as OLS regression of returns on the firm-specific explanatory variables and a constant.

The return of any stock can be expressed as:

$$r_i(n) = \alpha + x_i(n)\beta + \epsilon_{in}$$

where

$x_i(n)$ is the K dimensional vector of explanatory variables for the nth stock in sector i

β is a K vector of parameters, common to all sectors

α is the intercept term for all sectors

ϵ_{in} is a disturbance term assumed to be normally distributed, with mean 0 and constant variance across sectors: $\epsilon_{in} \sim N(0, \sigma^2)$.

5.5.2 Homoskedasticity: OLS on Firm-Specific Variables and Sector Dummy Variables

Here, we constrain the parameters as follows: $\beta_i = \beta_j$ across all sectors i,j, and $\theta_i = 0$ for all sectors i. The intercept term is allowed to vary between sectors. Then the return of any stock can be expressed as:

$$r_i(n) = \alpha_i + x_i(n)\beta + \epsilon_{in}$$

where

$x_i(n)$ is the K dimensional vector of explanatory variables for the nth stock in sector i ,

β is a K vector of parameters, common to all sectors,

α_i is the intercept term specific to sector i ,

ϵ_{in} is a disturbance term assumed to be normally distributed, with mean 0 and constant variance across sectors: $\epsilon_{in} \sim N(0, \sigma^2)$.

5.5.3 Groupwise Heteroskedasticity

We constrain $\alpha_i = \alpha_j$ for all sectors i,j, and $\theta_i = 0$ for all sectors i. Here, every sector is allowed its own variance.

Then the return of any stock can be expressed as:

$$r_i(n) = \alpha + x_i(n)\beta + \epsilon_{in}$$

where

$x_i(n)$ is the K dimensional vector of explanatory variables for the nth stock in sector i ,

β is a K vector of parameters, common to all sectors,

α_i is the intercept term specific to sector i ,

ϵ_{in} is a disturbance term assumed to be normally distributed, with mean 0 and variance σ_i^2unique to each sector i, so that $\epsilon_{in} \sim N(0, \sigma_i^2)$.

5.5.4 General Model

Here, we constrain the intercept α and the parameter vector β to be the same for all sectors. We also force θ to be the same for all sectors. Each sector i has a unique variance parameter γ_i. Then the return of any stock can be expressed as:

$$r_i(n) = \alpha_i + x_i(n)\beta + \epsilon_{in}$$

where

$x_i(n)$ is the K dimensional vector of explanatory variables for the nth stock in sector i ,
β is a K vector of parameters, common to all sectors,
α_i is the intercept term sector i,
ϵ_{in} is a disturbance term assumed to be normally distributed, with mean 0 and variance $\sigma_j^2(n)$ such that

$$\sigma_i^2(n) = exp(\gamma_i + x_i(n)\theta)$$

where

γ_i is a constant term specific to sector i.
θ is a K vector of parameters common to all sectors.

In the next chapter we estimate the above the OLS and GWH models for 88 months of data, and report the average of certain paramater values, and the average of measures of fit. We only estimate the general model for one month of data due to computational difficulties associated with the maximum likelihood estimation of the parameters.

Chapter 6

Evaluation of Probabilistic Models of Stock Returns

In this section we present the results obtained by estimating, for all datasets, the OLS model with and without sector dummy variables, and the GWH model. The more general form of the model requires estimation by maximum likelihood, and we did not have the computing facilities to efficiently run the routine for every dataset. We therefore only present the results from one month of data, and compare the results to the other models estimated for that particular month.

In all the above models, we use a set of three firm-specific variables: market equity of stock (ME), earnings to price ratio (E/P), book value over capitalization (BE/ME). We chose these explanatory variables based on the results by Fama and French (1992), who find these variables are significant in explaning cross-sections of stock returns in the American stock market.

6.1 OLS: without, and with, Sector Dummy Variables

The next two tables summarize the results of 90 OLS regressions of monthly returns on the three firm-specific variables and a constant. The mean number of observations per dataset was 1492, with a standard deviation of 442. We provide the mean, standard deviation, and pseudo-t-statistic of the estimated paramater values. These summary statistics are calculated over the 90 regressions. The pseudo-t-statistics are calculated by dividing the average parameter value by the estimated standard deviation of the average. The estimated standard deviation of the average is of course the sample standard deviation divided by the square-root of the number of observations, in this case 90. The pseudo-t-statistic of a parameter can be used to - heuristically, since these are not really t–statistics - test the hypothesis that on average, across all months, the variables associated with that parameter has no effect on average return. This procedure was introduced by Fama and MacBeth (1973), and used again by Fama and French (1992). This assumes that the estimated parameters are nearly independent from month to month. To verify this assumption, autocorrelations were calculated and are also provided below.

Parameter	Mean	Std.	t-statistic	Autocor.
Intercept	0.4284	6.2969	0.65	0.0760
ME	-15.1038	27.8262	-5.15	0.0208
E/P	-8.7031	54.9690	-1.50	0.2960
BE/ME	-10.4482	69.6006	-1.42	0.1437

Table 6.1: Average of Parameters estimated by OLS, t-statistics, and Autocorrelation Values across 90 datasets

As in Fama and French, and consistent with previous findings, it seems that the size (ME) variable has a negative parameter on average, implying that higher capitalization stocks have lower returns. Furthermore, the t-statistic (-5.15) is significant. Unlike Fama and French, however, we find that the E/P and BE/ME parameters have negative signs, and are insignificant. They find those parameters have positive signs and are significant. This may be due to the fact that we use Japanese data, as well as to the fact that our estimation period is shorter - they use monthly data for the U.S. stock market, from July 1963 to December. Perhaps more relevant, Chan et al. (1991) document that for the Japanese stock market, ME is insignificantly negative, E/P is insignificantly negative, and BE/ME is significantly positive. They note, that the signs and significance of the parameters are highly dependent on the model formulation, and on the variable definitions. They use a seemingly unrelated regression (SUR) framework across a set of monthly data, and include cash flow (C/P) as another variable in their regression. Our results are therefore not directly comparable.

We next report a summary of various measures of fit associated with our model.

The R^2 and Adjusted-R^2 are low for almost every month. This says that our three explanatory variables are not doing much in terms of explaining the cross-sectional variation in stock monthly returns. We cannot compare these last figures with other studies, since it seems to be

Measure of Fit	Mean	Std.
R^2	0.0524	0.0504
Adjusted-R^2	0.0499	0.0506
Mean-AIC	7.0822	0.4037

Table 6.2: Average and Standard Deviation of Measures of fit for OLS across 90 datasets

the norm not to report such measures of fit, probably because they are low - again, it's useful to remember that it would be preposterous to expect that any combination of firm-specific variables such as these would explain much of the variation in returns in any month, given the inherent difficulty of explaining why returns vary as they do .

When we include sector dummy variables, the sign of the parameter ME changes, and the parameters E/P and BE/ME become significant.

Parameter	Mean	Std.	t-statistic	Autocor.
ME	0.0158	1.2955	0.12	-0.0272
E/P	-10.0540	24.5713	-3.88	0.2919
BE/ME	-1.0126	11.5197	-6.32	0.2664

Table 6.3: Average of Parameters estimated by OLS (Including Unique Sector Intercept), t-statistics, and Autocorrelation Values across 90 datasets

We only keep sectors that have at least two stocks in them, so we reduce the size of our datasets, slightly - mean number of observations is 1489, with a standard deviation of 443. Not every month has the same nmber of stocks, or the same number of sectors, since sectors with a small number of stocks may not be represented when those stocks are absent. The mean number of sectors, calculated across time, is 50, with a standard deviation of 5.

The measures of fit below show that including the industry sector dummy variables does improve the fit of the model, even when penalizing for more parameters - higher Adjusted-R^2 and lower mean AIC.

Measure of Fit	Mean	Std.
R^2	0.1718	0.0728
Adjusted-R^2	0.1355	0.0700
Mean-AIC	7.0242	0.3799

Table 6.4: Average and Standard Deviation of Measures of fit for OLS across 90 datasets

6.2 GWH Model

Here, we obtain our results by using an iterative least-squares algorithm. The steps of the algorithm are:

1. Estimate the residual variances for every industry sector i, using the formula $\sigma_i = e_i'e_i/n_i$, where e_i is the residual variance vector for sector i, and n_i is the number of observations in sector i.

2. Compute β according to $\beta = \left[\sum_i \frac{1}{\sigma_i}(X_i'X_i)\right]^{-1} \sum_i \frac{1}{\sigma_i}(\mathrm{X}_i'y_i)$.

3. If β has not converged, go to step 2.

This provides us with maximum-likelihood estimates of the parameters, but is much quicker than using standard nonlinear optimization[1]. We work with 88 datasets, having had to eliminate some datasets which caused problems with the algorithm. The datasets which we eliminate are 04/89, 01/92, and 05/95. The mean number of observations across all estimations is 1505, with a standard deviation of 428. The results, parameter statistics and measures of fit, are in the two tables below. Note that R^2 measures are not reported here, because of the lack of valuable interpretation outside of the OLS regression framework. We rely on the mean AIC as an indicator of how well we are doing.

Parameter	Mean	Std.	t-statistic	Autocor.
Intercept	0.0771	33.0700	0.02	0.0943
ME	-0.0136	1.2285	-0.10	-0.1041
E/P	-8.2542	19.6495	-3.94	0.2919
BE/ME	-0.6212	1.3066	-4.45	0.2664

Table 6.5: Average of Parameters estimated by the GWH Model, t-statistics, and Autocorrelation Values across 88 datasets

We can compare the GWH and OLS - without industry sector dummy variables - by using an appropriate χ^2 test. Let H_0: return residuals are homoskedastic, and H_1: return residuals are heteroskedastic. Then, we use the fact that $-2(L_0 - L_1)$ has a χ^2 distribution with degrees

[1] See Greene, p. 369.

Measure of Fit	Mean	Std.
Mean-AIC	6.9727	0.3941

Table 6.6: Average and Standard Deviation of Mean-AIC : GWH Model across 88 datasets

of freedom equal to the number of sector variances which we estimate minus one. Of course, L_0 is the likelihood of the OLS model, and L_1 is the likelihood of the GWH model. Note also the following relationships: $-2(L_0 - L_1) = n\log(s^2) - \sum_i(n_i \log(s_i^2))$ We run this test for 88 datasets, and reject H_0 in favor of H_1 in all but two cases. Both the AIC values and the χ^2-test described above lend support to the hypothesis of heteroskedasticity. If this hypothesis is accepted then incorporating the inherent heteroskedasticity of cross-sectional returns into our parameter estimation will allow us to obtain more efficient estimates - i.e., estimates with smaller standard deviation. We would trust the parameter estimates obtained by incorporating the heteroskedasticity of returns more than we would trust the results obtained from using OLS. It appears then that size is not a significant variable, but that E/P and BE/ME are. Furthermore, these last two parameters are negatively correlated with return, other things being equal. Note however that these results must be taken with a grain of salt, since the parameter estimates seem to be so sensitive to model specification. The most consistent fact that we uncover across the three model specifications is that the E/P and BE/ME parameters have negative signs, implying that high E/P values or high BE/ME values are associated with lower returns on average, which is counter intuitive. The ME parameter shifted in sign, going from negative to positive between OLS without and OLS with industry sectors. It is, however, insignificant in OLS with sector dummies and the GWH specification.

6.3 Implementing the General Model of Stock Returns

The general model described earlier does not lend itself to efficient calculation. Its parameters are estimated by maximum likelihood, but this involves the optimization of a nonlinear likelihood function. We therefore restricted our attention to one month of data, January 1993. We used the Splus optimizer function `ms` to carry out the calculation.

For easy reference, we recapitulate the structure of the general model. Here, the return of any stock can be expressed as:

$$r_i(n) = \alpha_i + x_i(n)\beta + \epsilon_{in}$$

where

$x_i(n)$ is the K dimensional vector of explanatory variables for the nth stock in sector i ,

β is a K vector of parameters, common to all sectors,

α_i is the intercept term for sector i,

ϵ_{in} is a disturbance term assumed to be normally distributed, with mean 0 and variance $\sigma_{\jmath}^2(n)$ such that

$$\sigma_i^2(n) = exp(\gamma_i + x_i(n)\theta)$$

where

γ_i is a constant term specific to sector i.

θ is a K vector of parameters common to all sectors.

Note that only sectors with more than 30 observations are used in order to guarantee the significance of the parameters and to increase computation speed by reducing the total number of parameters to be estimated. Below, we list the sectors that we use, and indicate the number of stocks that they contain.

Sector Number	4	8	9	12	16	17	21	24	25	28	36
Number of Stocks	80	51	86	42	80	39	79	34	71	51	137
Sector Number	41	43	44	49	54	57	59	61	63	64	
Number of Stocks	58	33	80	36	39	49	34	55	142	44	

Table 6.7: Sectors and Number of Stocks They Contain

The following tables give the results of the estimation.

First, notice that the sector dependent parameter values, α_i and $\gamma_\imath$, are of the same magnitude across sectors. This seems to indicate that their may be some advantage in combining

Parameter	ME	E/P	B/M	α_4	α_8	α_9	α_{12}	α_{16}	α_{17}	α_{21}
Estimate	0.86	-6.81	-0.98	-24.51	-23.35	-22.98	-21.50	-22.02	-23.46	-24.94
t-statistic	7.06	-4.86	-2.75	-7.62	-7.63	-7.52	-6.86	-7.11	-7.73	-8.32
	α_{24}	α_{25}	α_{28}	α_{36}	α_{41}	α_{43}	α_{44}	α_{49}	α_{54}	α_{57}
	-22.94	-22.79	-25.84	-22.93	-21.83	-22.77	-25.62	-23.37	-22.81	-23.65
	-7.39	-5.79	-8.62	-7.68	-6.86	-7.40	-8.44	-7.40	-6.93	-7.90
	α_{59}	α_{61}	α_{63}	α_{64}						
	-23.65	-23.99	-23.85	-23.34						
	-7.51	-7.98	-7.91	-7.15						

Table 6.8: General Model: parameters affecting mean

Parameter	ME	E/P	B/M	γ_4	γ_8	γ_9	γ_{12}	γ_{16}	γ_{17}	γ_{21}
Estimate	0.09	-1.83	-0.38	5.06	5.21	5.48	5.34	5.58	5.46	5.75
t-statistic	-2.47	-1.49	-4.80	5.37	5.68	6.38	5.77	6.12	6.18	6.93
	γ_{24}	γ_{25}	γ_{28}	γ_{36}	γ_{41}	γ_{43}	γ_{44}	γ_{49}	γ_{54}	γ_{57}
	4.85	6.46	5.40	5.55	5.38	5.14	4.55	4.98	5.88	5.23
	5.31	7.58	6.14	6.26	5.79	5.76	4.99	5.32	6.16	5.95
	γ_{59}	γ_{61}	γ_{63}	γ_{64}						
	5.24	5.04	5.22	5.45						
	5.57	5.61	5.88	5.97						

Table 6.9: General Model: parameters affecting variance

sectors whose parameters are most alike. This idea will be examined further in the following chapters on clustering. We will see that combining sectors. based on their closeness in terms of their sector specific parameters offers some improvement in the measure of fit of the model. Also, notice that the t-statistics of these parameters indicate that they are all individually significant. The other parameters are also significant, except for the ME parameter affecting the variance, which has a t-statistic of -1.49 .

For the same dataset, we also ran the OLS with and without dummy variables, and the GWH model. The results are below, where we reproduced the general model parameters that were also in the other models. Notice that the estimates of the ME, E/P, and B/M variables affecting the mean are within the same magnitude for all models, including the general model. They also have the same signs. Of course, for the OLS with dummy variables, and the GWH models, other parameters were estimated, but they are not reported in the following table, not being shared by all models. The AIC is lowest for the general model. indicating that for this dataset, the general model is superior in characterizing the mechanisms of cross-sectional return behavior.

Parameter	OLS	OLS and dummy	GWH	General Model
ME	0.88 (7.13)	0.87 (6.45)	0.89 (73.46)	0.86 (7.06)
E/P	-8.46 (-2.5)	-6.81 (-2.00)	-7.82 (-0.92)	-6.81 (-4.86)
B/M	-0.79 (-2.39)	-0.97 (-2.75)	-0.80 (-9.17)	-0.98 (-2.75)
R^2	0.062	0.091	NA	NA
Adjusted-R^2	0.059	0.0743	NA	NA
AIC	8392.81	8392.89	8262.79	8223.82

Table 6.10: Parameters Common to all Models, estimated for January 1993, and Measures of Fit

Chapter 7

Description of Clustering Algorithms

In this chapter, we motivate the application of clustering techniques to our initial set of industry sectors. We then describe two types of clustering algorithm, the hierarchical and the non-hierarchical. We end up applying the non-hierachical K-means algorithm to our problem..

At this point, some terminology must be defined to clearly present the ideas behind clustering. Both types of algorithm use as input a set of *industry sector variables* - variables that characterize each sector. An example of such a set of variables is the average market equity (ME), the average earnings to price ratio (E/P), the average book to market ratio (B/M), the average return, the average standard deviation of return, over all stocks in any given sector. Another set of industry variables could be industry specific parameters from the models presented in earlier chapters. We call *cluster* the result of merging one or more sectors. A single sector is a cluster. Of course, a cluster is also a sector that contains all the stocks in the sectors that were merged into it. Then, given a set of industry sector variables, our clustering algorithms determine what clusters are closest to each other, according to a specified metric in the space of industry sector variables. Notice that we have to define the distance between sectors or clusters. In the hierarchical case, the two closest clusters or sectors are merged at every iteration of the clustering algorithm. In the non-hierachical case, a number of final clusters is chosen in advance, and the clustering algorithm merges the initial sectors so as to form homogeneous clusters.

7.1 Motivation for Clustering Sectors

The following supplies the motivation and logic behind the clustering algorithms that we present, and propose to implement, and outlines their structure. In this research, the clustering of stocks is intended to increase the predictive power of our model, as measured by its AIC. Remember that the expression for the AIC is

$$-2(\text{Loglikelihood of Model}) + 2(\text{Number of Parameters in the Model})$$

Our models describe the probabilistic behavior of stock returns in terms of firm-specific explanatory variables, and industry sector dummy variables. It may be possible to increase the predictive power of the model - or equivalently, to decrease its AIC - by merging sectors, since that would decrease the number of sector specific parameters to estimate. On the other hand, the AIC will improve only if the likelihood of the model does not decrease too much. Consider merging two sectors; in the context of the GWH model. Then the number of parameters to estimate is decreased, by two. To see this, remember that each sector has unique mean and variance terms. Hence,if the loglikelihood of our model does not decrease significantly in the process of merging, the AIC will decrease, improving the quality of the model.

On the other hand, splitting sectors, in the case where the cross-sectional behavior of its stocks is heterogeneous, might also decrease the AIC. Refering again to the definition of the AIC, suppose that splitting two sectors increases the loglikelihood significantly. Then, even though the number of parameters to estimate is increased, the AIC will decrease. This last point, however, is outside the scope of this research, and we consider the more limited case where we allow merging of sectors only.

Remember that we decide to start with a given industry classification provided by a financial institution. We can therefore consider merging any of the given sectors, to obtain a new industry classification. The classification matrix C which was described in the section where the probabilistic model was developed, contains columns which correspond to sectors. Each column i is a dummy variable vector, with a 1 in row n if stock n is in sector i, and a 0 otherwise. In terms of the classification matrix, merging of any pair of sectors, say i and j consists in adding the two columns i and j to obtain a new column. This column corresponds to a new sector, with all the stocks in i and j, and can be added to C. Columns i and j are then deleted from C.

7.2 Hierarchical Clustering Algorithm

We cannot try all possible combinations of sectors. Take the Vestek industry classification, which has more than 60 sectors. There are just too many cases to consider, if we were to try every possible combination. Instead, we use a dynamic approach to attack the problem. We start with a given set of sectors, and its corresponding sector variable space. We estimate the model that uses the corresponding industry classification, and record its AIC. We start with as many clusters as there are original sectors. We then merge the two clusters which are "closest" according to some distance measure in the sector variable space. We repeat the process, until all sectors have been merged into one final cluster, at each iteration reducing the number of clusters by one. Note that we still need to define what we mean by the distance between two sectors. This can be done in several ways. We present three of the most common choices. In *single linkage*, the distance between two clusters is the minimum distance between any sector in the first cluster and any sector in the second. In *complete linkage*, the distance between two clusters is the maximum distance between any sector in the first cluster and any sector in the second. In *average linkage*, the distance between two clusters is the average of all distances between any sector in the first cluster and any sectors in the second.

There are therefore four steps in our clustering algorithm:

Step 1: Calculate the distance matrix between all original clusters.

Step 2: Merge the two clusters that are closest according to a chosen distance metric.

Step 3: Update the distance matrix, which has one less element.

Step 2: If there is more than one cluster left, go back to Step 1.

7.3 Non-hierachical Clustering Algorithms

The method proposed here is also called the K-means algorithms. It requires that the number of final groups of sectors to be obtained be specified. Given we have an initial partitioning of our sectors into K groups, we then follow the simple algorithm. We calculate the mean of each variable across the sectors in each group. We then obtain a vector of variable means, called a centroid, for each group. We then iterate between the following steps:

Step 1: Choose a group. Pick a sector, and assign it to the group with the nearest centroid. If the nearest centroid is the one in which the sector already is, it will not change groups. Otherwise, update both the centroid that looses and the centroid that receives an element.

Step2: Repeat Step 1 until no more no sector can change groups.

Appendix B goes into the mathematical formulation of the K-means algorithm. Another interesting approach to clustering involves fuzzy logic objective function, and is also described in Appendix B.

7.4 Distance Measures

At this point we examine several distance measures, applicable to our problem. We only select a few of them for actual implementation.

7.4.1 A Simple Distance Measure : the Covariance or Correlation Matrices of Industry Returns

We use the same method as Farrel (1974), but we apply it to industry sectors instead of industry stocks. Farrel calculates the correlation matrix for stock returns, and then uses a hierarchical clustering approach to merge stocks, or groups of stocks, which have the highest correlation at each step of the algorithm. The equivalent distance measure that fits our purposes is the correlation coefficient between sector returns. Sector returns are defined to be the average of the returns in the sector. Therefore, we assume that the distance between two sectors is simply the correlation between them. In the next chapter we implement this algorithm for the GWH model, which can be estimated fast for the whole eight years of data. Another approach, working with the covariance matrix of sector returns, is to rank the variances, and combine sectors within the same variance decile. This method would yield ten sectors from the original of about 60, and this new condensed industry classification could be combined with the first. The problem with distance measures proposed here is that they rely on the time series behavior of returns, as opposed to their cross-sectional return. Hence promising results are not necessarily expected, but the distance measures were presented anyway, as examples of distance measures that could be appropriate in another model setting, where the time-series evolution of returns is not ignored.

7.4.2 Distance Measure Based on Firm-Specific Variables, Intra-Sector Mean Return and Standard Deviation.

Each sector, at each time period, can be associated with the following five variables: the mean ME, E/P and B/M *of the stocks which it includes,* and the mean and standard deviation of the returns of these stocks. These variables were introduced earlier as *sector variables*, as they characterize each sector at each time period. We then take the average of these sector variables across all time periods, to obtain the variables that were introduced earlier as *average sector variables.* We standardize these variables - substract the mean calculated across all sectors, and divide by the standard deviation across all sectors. We then take as a measure of distance between sectors their Euclidean distance in the five-dimensional space of their characteristic variables. That is

$$d(i,j) = \sqrt{(r_i - r_j)^2 + (\sigma_i - \sigma_j)^2 + (\mathrm{ME}_i - \mathrm{ME}_j)^2 + (\mathrm{E/P}_i - \mathrm{E/P}_j)^2 + (\mathrm{B/M}_i - \mathrm{B/M}_j)^2}$$

where

i, j are sectors, $1 \leq i, j \leq K$, where K is the number of sectors,

r_i, σ_i, ME_i, $\mathrm{E/P}_i$, $\mathrm{B/M}_i$, are the return, standard deviation, market equity, earnings to price ratio, book to market ratio, of sector i, averaged over all time periods, and standardized over all sectors.

Note that we could also just consider any subcollection of the above five variables, in our definition of a distance measure. For example. if we want to avoid using the firm specific variables completely, and just focus on the characteristics of sector returns, we could define the distance between two sectors i and j to be:

$$d(i,j) = \sqrt{(r_i - r_j)^2 + (\sigma_i - \sigma_j)^2}$$

where

i, j are sectors, $1 \leq i, j \leq K$, where K is the number of sectors,

r_i, σ_i, are the return and standard deviation of sector i, averaged over all time periods, and standardized over all sectors.

7.4.3 Distance Measure in Terms of the Parameters in the General Model

We now describe three possible distance measures to use in the above algorithms, all based on the comparison of the parameters from the general model. All distance measures assume that sectors are defined solely in terms of the parameters of the probabilistic model. Hence, we do not use any outside information about the sectors, such as their economic definition.

Remember the expression for the mean and variance of returns, as defined in our model. The expression for the mean of sector i is given by

$$\mu_i(n) = \alpha_i + x_i(n)\beta$$

where

$x_i(n)$ is the K vector of explanatory variables of the nth stock in sector i,

α_i intercept term for group i,

β is a K vector of parameters common to all sectors.

The expression for the variance term of the nth stock in sector i is

$$\sigma_i(n) = \exp(\gamma_i + x_i(n)\theta)$$

where

$x_i(n)$ is the K vector of explanatory variables of the nth stock in sector i,

γ_i is a constant term specific to sector i.

θ is a K vector of parameters, to be estimated.

Given an industry classification, it follows that each sector i can be defined in terms of its two paramaters α_i and γ_i. Consider a pair of sectors i and j, where $0 \leq i < j \leq J$, J being the total number of sectors in the industry classification. We define their difference vector d_{ij} as

$$d_{ij} = \begin{bmatrix} (\alpha_i - \alpha_j) \\ (\gamma_i - \gamma_j) \end{bmatrix}.$$

The three distance measures which we propose are:

1. Euclidean Norm $= \sqrt{d_{ij}' d_{i,j}} = \sqrt{(\alpha_i - \alpha_j)^2 + (\gamma_i - \gamma_j)^2}$. The Euclidean distance measure assumes that differences in the sector mean parameter α are as important as differences in the variance parameter γ.

2. Weighted distance $= \lambda(\alpha_i - \alpha_j)^2 + (1 - \lambda)(\gamma_i - \gamma_j)^2$, where $0 \leq \lambda \leq 1$. In the extreme case where $\lambda = 1$, only differences in the α terms are accounted for. We are then saying that only differences in the mean matter. In the other extreme case, $\lambda = 0$, and we are saying that differences in the mean don't count. For other values of λ, we are weighting the α and γ terms differently. This could be especially useful if the α and γ terms are on different scales.

3. Statistical distance $= d_{ij}' COV(d_{ij})^{-1} d_{ij}$. Notice that the statistical distance measure is equivalent to a Wald test for the hypotheses that $(\alpha_\imath - \alpha_j) = 0 \Leftrightarrow \alpha_i = \alpha_j$ and $\gamma_i - \gamma_j = 0 \Leftrightarrow \gamma_i = \gamma_j$. An estimate of $COV(d_{ij})$ is obtained from the covariance matrix of the model parameters. Remember that the parameters are estimated by maximum-likelihood, so that we can obtain an efficient estimate of the covariance matrix.

4. Standardized Norm: here, we standardize all parameters by substracting their inter-sector average, and dividing by their inter-sector standard deviation. Specifically, we do the following transformations:

$$\alpha_i \longrightarrow \alpha_i^* = \frac{(\alpha_i - \frac{(\sum_\imath \alpha_\imath)}{J})}{\frac{(\sum_\imath \alpha_i^2)}{J} - (\frac{(\sum_\imath \alpha_i)}{J})^2},$$

$$\gamma_i \longrightarrow \gamma_i^* = \frac{(\gamma_i - \frac{(\sum_\imath \gamma_\imath)}{J})}{\frac{(\sum_i \gamma_\imath^2)}{J} - (\frac{(\sum_\imath \gamma_\imath)}{J})^2},$$

where J is the total number of sectors.

We then define a standardized difference vector, d^*_{ij}, as

$$d^*_{ij} = \begin{bmatrix} (\alpha^*_i - \alpha^*_j) \\ (\gamma^*_i - \gamma^*_j) \end{bmatrix}.$$

and the Standardized Euclidean Norm $= \sqrt{d^{*\prime}_{ij} d^*_{ij}} = \sqrt{(\alpha^*_i - \alpha^*_j)^2 + (\gamma^*_i - \gamma^*_j)^2}$. This is also just the Euclidean distance, applied to the standardized parameters. A weighted distance could also be defined in terms of the standardized parameters.

It should be pointed out that the statistical distance measure has an inherent weakness. This comes from the fact that if a sector i has parameters α_i and γ_i which have very high variances, then the statistical distance measures between i and any other sector will be very low. This does not reflect our understanding of "closeness" between two sectors, and is counter-intuitive. Furthermore, merging the aforementionned sector i with another sector will create a sector with parameters having high variance. Hence, the merged sectors will act as a magnet for other sectors and, similar to a snow-ball effect, after several iterations each new iteration merges one of the original sectors to a big high variance super sector. This effect was noticed in practice, and so we do not further deal with this particular distance metric.

Of the other metrics mentionned aboved, the plain Euclidean distance does not take into account the scales of each parameter, and will completely neglect parameters that on average have small values. Again, this may not reflect our understanding of closeness, since these parameters may just take values across sectors on a different scale. In the case of the weighted distance measure, the weight λ is not specified, and could be determined empirically.

In the following empirical work, we have chosen to focus exclusively on the most logical between the above distance measures, the standardized Euclidean norm.

Chapter 8

Clustering Algorithm Results

In this chapter, we implement the K-means clustering algorithm described above. We choose to create $K=$ 40, 30, 10 clusters from the original 55 groups in our first application of the K-means algorithm, based on the Euclidean distance between industry variables. In our second application of the K-means algorithm, where we cluster based on the parameters of the general model, we choose K=15, 10, 5 clusters from the original 21 groups - remember that in the general model, we eliminate some small groups from the original 55, and are left with 21 groups. Using the clusters formed by the K-means clustering algorithm, we run the GWH or General Model again using the new industry classification. Again, we obtain parameter estimates and measures of fit. We compare the results we obtain to the results obtained with the original industry classification.

. Before we present the clustering algorithm results, we look at a simple way of grouping sectors, by ranking sectors based on their time series standard deviation, and grouping the 10 deciles of this ranking. This simple approach is shown below.

8.1 Grouping Industry Sectors by Examining the Time Series Standard Deviation of Sector Returns

Here we calculate the standard deviation of sector returns, over all 88 time periods used to run the GWH model with the original classification. Remember that a sector return is defined as the average of the returns of all stocks included in the sector. We rank standard deviations into ten groups, from lowest to highest decile. We then group sectors in the same decile, and use this new condensed industry classification as an input to the GWII model. Remember that the GWH model assumes that returns are a function of one constant and three explanatory variables, and that the standard deviation of the returns are sector-specific, but are constant within a sector. We obtain the following results.

Parameter	Mean	Std.	t-statistic	Autocor.
Intercept	0.3955	34.2247	1.02	0.2923
ME	-0.0217	1.2714	-1.50	0.1626
E/P	-9.2650	23.5713	-34.59	0.3601
BE/ME	-0.7438	1.3229	-49.48	0.3169

Table 8.1: Average of Parameters estimated by the GWH Model (new industry classification after using the kmeans clustering algorithm to btain 10 groups), t-statistics, and Autocorrelation Values across 90 datasets

Measure of Fit	Mean	Std.
Mean-AIC	7.0325	0.3927

Table 8.2: Average of Parameters estimated by OLS, t-statistics, and Autocorrelation Values across 90 datasets

Remember that the GWH model, when applied to the original industry classification, over the entire 88 datasets which we used, yielded an average mean AIC of 6.9727. We are therefore increasing the AIC by combining sectors with the preceding algorithm. There is therefore loss of information in the process, and we are not doing as well as with our original industry classification. This was to be expected given the nature of the grouping process here. Indeed, the distance measure that we used, the time series standard deviation of sector returns, is not associated with intra-sector return standard deviation. But intra-sector standard deviation is the quantity of interest in the GWH framework, since we are modelling the standard deviation of returns within each sector at each time period, rather than the standard deviation across time of sector returns.

8.2 Clustering Based on the Comparison of the Sector Specific Variables Return, Standard Deviation of Return, ME, E/P, BE/ME.

Here, we implement the non-hierachical K-means clustering algorithms. The first step, as when using any clustering algorithm, is to define a distance measure. We use the distance measures based on the sector variables. These distance measures were defined in the previous chapter, and are reproduce here for convenience. The distance measure based only on the standardized mean return and standard deviation is

$$d(i,j) = \sqrt{(r_i - r_j)^2 + (\sigma_i - \sigma_j)^2}$$

where

i, j are sectors, $1 \leq i, j \leq K$, where K is the number of sectors,

r_i, σ_i, are the return and standard deviation of sector i, averaged over all time periods, and standardized over all sectors.

The more complete distance measure that incorporates our knowledge about the firm-specific variables included in the sectors under consideration is

$$d(i,j) = \sqrt{(r_i - r_j)^2 + (\sigma_i - \sigma_j)^2 + (\mathrm{ME}_i - \mathrm{ME}_j)^2 + (\mathrm{E/P}_i - \mathrm{E/P}_j)^2 + (\mathrm{B/M}_i - \mathrm{B/M}_j)^2}$$

where

i, j are sectors, $1 \leq i, j \leq K$, where K is the number of sectors,

r_i, σ_i, ME_i, $\mathrm{E/P}_i$, $\mathrm{B/M}_i$, are the return, standard deviation, market equity, earnings to price ratio, book to market ratio, of sector i, averaged over all time periods, and standardized over all sectors.

We call the first distance measure d_1 and the second distance measure d_2 . We start with 54 Vestek industry sectors - we only keep those for which we could calculate all the variables used to measure distance, namely r_i, σ_i, ME_i, $\mathrm{E/P}_i$, $\mathrm{B/M}_i$ as defined above. For both distance measures, we implement the K-means algorithm and form a number of groups K out of the original 54 sectors. We initialize the algorithm using the output of the non-hierachical clustering algorithm that uses the same distance measure. This output can be represented by a dendogram, which is basically a tree whose nodes represent mergings. We do not show the dendograms here. The dendogram can be cut to obtain K groups. This dendogram-cuting is done without any optimization orientation, and is used only to obtain an initial guess as to what the K groups would be.

The following four tables give the results of the K-means clustering algorithm. We show for each sector in the original Vestek industry classification, its mapping into the lower dimensional K -group classification, for $K = 10,30,40$, and for both distance measures mentionned above. We also show the results in the principal components space of the matrix of standardized average sector variables.

We use each of the classifications obtained above to run the GWH model on 88 months of data. For each of the classifications, we report the mean, standard deviation, t-statistics and autocorrelation of the parameters over the 88 months. We also report the mean AIC value averaged over all months.

Parameter	Mean	Std.	t-statistic	Autocor.
Intercept	1.3098	3.6583	0.36	0.
ME	-0.0572	0.1360	-0.4208	0.
E/P	-9.8121	2.4189	-4.05	0.
BE/ME	-0.8120	0.1443	-5.62	0.

Table 8.7: Average of Parameters estimated by the GWH Model (new industry classification after using the kmeans clustering algorithm to btain 10 groups), t-statistics, and Autocorrelation Values across 90 datasets

Measure of Fit	Mean	Std.
Mean-AIC	7.0379	0.3987

Table 8.8: Average of Mean AIC of the GWH Model (new industry classification after using the K-means clustering algorithm to obtain 10 groups, d=d1)

Vestek Industry Code	$K=10,d=d_1$	$K=30,d_1$	$K=40,d=d_1$
1	3	10	8
2	3	11	9
3	2	12	10
4	4	13	11
5	3	3	6
6	NA	NA	NA
7	6	19	12
8	1	1	2
9	6	5	4
10	6	14	13
11	1	2	14
12	1	1	2
13	NA	NA	NA
14	3	3	6
15	3	6	15
16	1	1	2
17	7	27	16
18	NA	NA	NA
19	3	15	17
20	1	8	1
21	1	1	2
22	7	7	18
23	NA	NA	NA
24	3	6	19
25	3	3	20
26	NA	NA	NA
27	5	16	21
28	1	1	7
29	NA	NA	NA
30	1	8	7
31	7	17	22
32	NA	NA	NA
33	NA	NA	NA
34	4	18	23
35	6	19	24

Table 8.3: Mapping of Vestek industry sectors 1 through 30 into lower dimensional classification given by K-means algorithm: distance is d1

Vestek Industry Code	K=10,d=d_2	K=30,d=d_2	K=40,d=d_2
1	1	10	10
2	5	11	11
3	6	12	12
4	4	13	13
5	1	1	5
6	NA	NA	NA
7	1	14	14
8	3	2	2
9	3	26	9
10	3	23	8
11	9	6	15
12	3	2	2
13	NA	NA	NA
14	1	1	5
15	1	15	16
16	3	2	2
17	3	3	17
18	NA	NA	NA
19	1	16	18
20	3	2	2
21	7	20	6
22	9	7	19
23	NA	NA	NA
24	1	5	20
25	1	23	21
26	NA	NA	NA
27	4	17	22
28	3	3	6
29	NA	NA	NA
30	7	20	23
31	7	18	24
32	NA	NA	NA
33	NA	NA	NA
34	4	19	25
35	3	2	26

Table 8.4: Mapping of Vestek industry sectors 1 through 30 into lower dimensional classification given by K-means algorithm: distance is d2

Vestek Industry Code	k=10,d=d_1	k=30,d=d_1	k=40,d=d_1
36	7	7	3
37	NA	NA	NA
38	2	20	25
39	1	1	2
40	7	21	26
41	1	8	1
42	NA	NA	NA
43	1	1	2
44	1	2	27
45	NA	NA	NA
46	1	5	4
47	6	22	28
48	NA	NA	NA
49	3	23	29
50	NA	NA	NA
51	1	5	30
52	3	14	31
53	1	8	1
54	1	9	5
55	8	24	32
56	2	25	33
57	1	1	7
58	8	26	34
59	1	8	1
60	NA	NA	NA
61	7	27	35
62	NA	NA	NA
63	1	9	5
64	6	4	36
65	5	28	37
66	1	1	3
67	9	29	38
68	6	4	39
69	10	30	40

Table 8.5: Mapping of Vestek industry sectors 36 through 69 into lower dimensional classification given by K-means algorithm: distance is d1

Vestek Industry Code	k=10,d=d_2	k=30,d=d_2	k=40,d=d_2
36	3	3	4
37	NA	NA	NA
38	6	21	27
39	3	2	2
40	7	22	28
41	3	4	3
42	NA	NA	NA
43	9	6	29
44	3	2	2
45	NA	NA	NA
46	3	26	9
47	5	8	1
48	NA	NA	NA
49	1	23	30
50	NA	NA	NA
51	3	4	8
52	1	5	31
53	3	4	3
54	3	4	3
55	5	8	1
56	6	24	32
57	9	6	4
58	8	25	33
59	3	26	34
60	NA	NA	NA
61	9	7	35
62	NA	NA	NA
63	3	3	36
64	5	9	7
65	2	27	37
66	9	28	38
67	2	29	39
68	5	9	7
69	10	30	40

Table 8.6: Mapping of Vestek industry sectors 36 through 69 into lower dimensional classification given by K-means algorithm: distance is d2

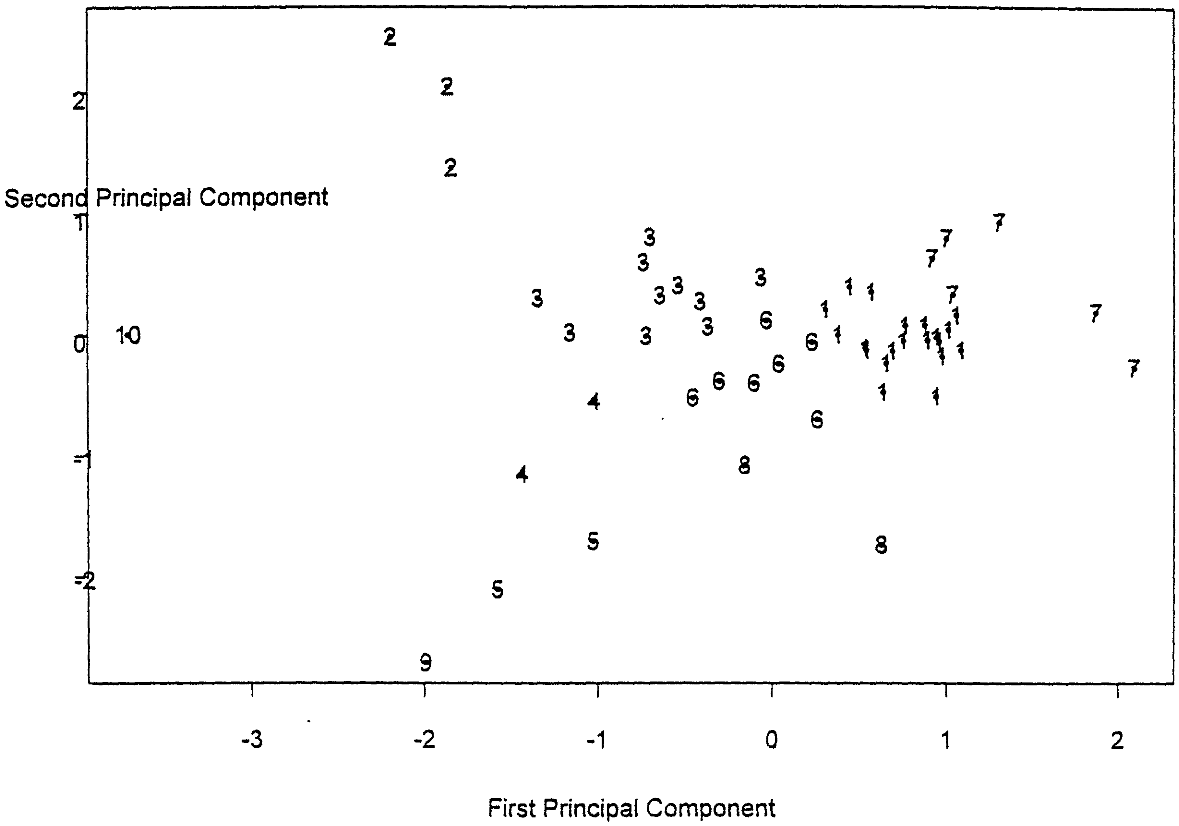

Figure 8-1: K-means Algorithm, K = 10, distance measure = d1

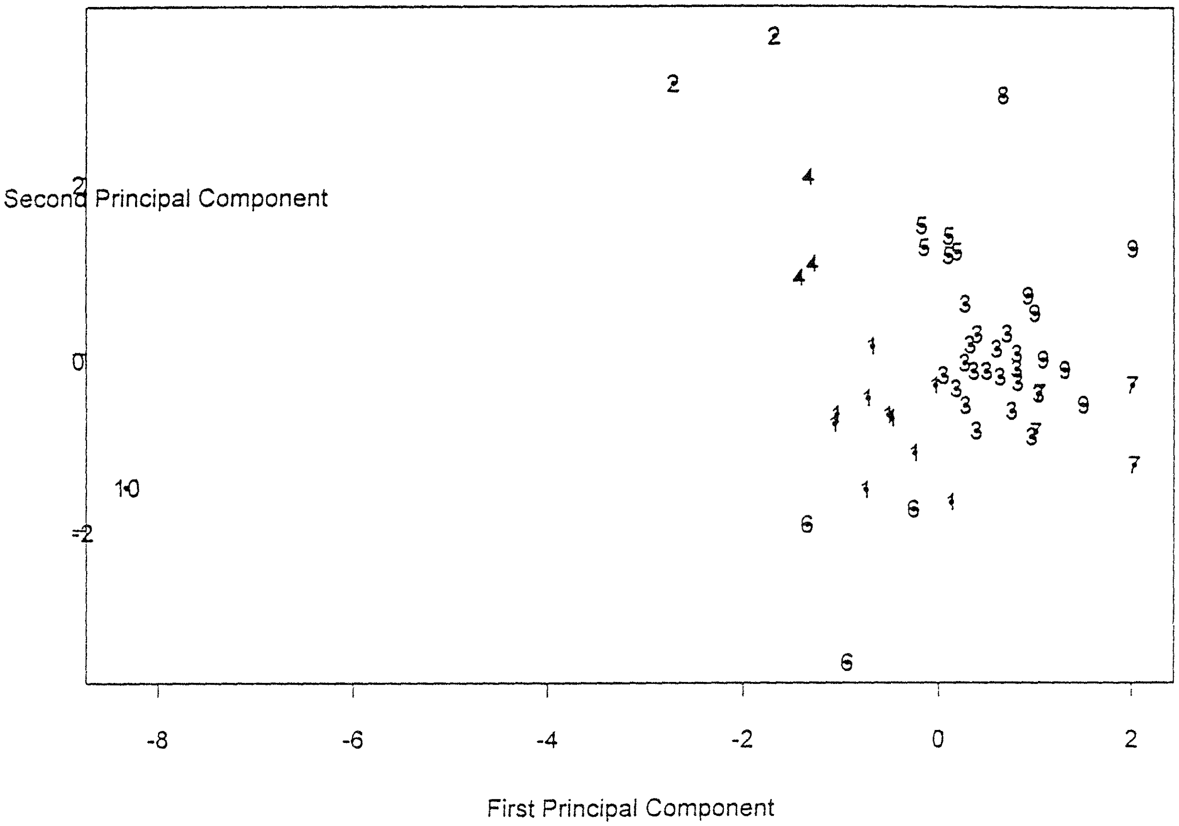

Figure 8-2: Result of K-means Algorithm, K = 10, distance measure = d2

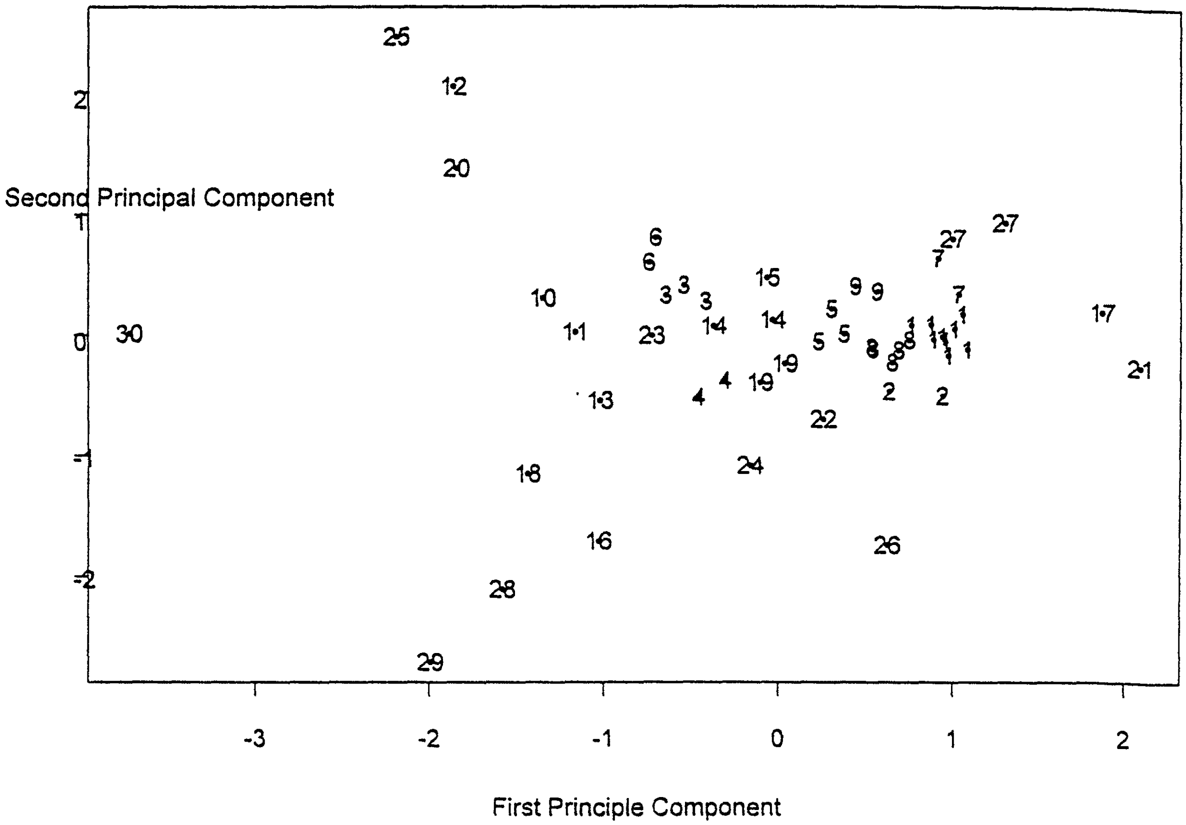

Figure 8-3: Result of K-means Algorithm, K = 30, distance measure = d1

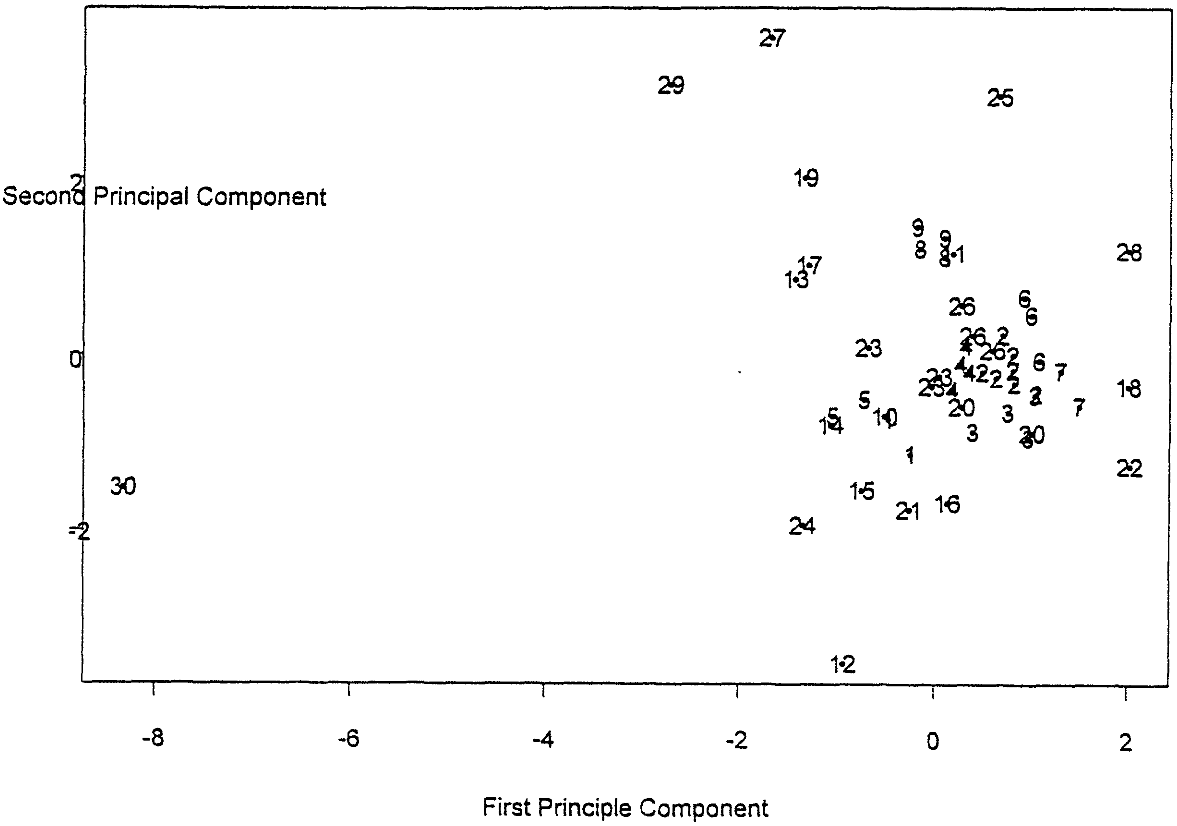

Figure 8-4: Result of K-means Algorithm, K = 30, distance measure = d2

Parameter	Mean	Std.	t-statistic	Autocor.
Intercept	1.2239	33.9968	0.3377	0.2923
ME	-0.0523	1.2636	-0.3885	0.1626
E/P	–2.2849	21.7312	-4.0080	0.3601
BE/ME	-0.7568	1.3430	-5.2862	0.3169

Table 8.9: Average of Parameters estimated by the GWH Model (new industry classification after using the kmeans clustering algorithm to btain 10 groups,d=d2), t-statistics, and Autocorrelation Values across 90 datasets

Measure of Fit	Mean	Std.
Mean-AIC	7.1796	0.4158

Table 8.10: Average of Mean AIC of GWH Model (new industry classification after using the kmeans clustering algorithm to btain 10 groups,d=d2)

Parameter	Mean	Std.	t-statistic	Autocor.
Intercept	0.9294	33.7707	0.2582	0.0958
ME	-0.0448	1.2524	-0.3355	0.0405
E/P	-9.6203	22.1131	-4.0811	0.0737
BE/ME	-0.7158	1.3240	-50715	0.1263

Table 8.11: Average of Parameters estimated by the GWH Model (new industry classification after using the kmeans clustering algorithm to obtain 30 groups, d=dl), t-statistics, and Autocorrelation Values across 88 datasets

Measure of Fit	Mean	Std.
Mean-AIC	7.0032	0.3955

Table 8.12: Average of Mean AIC of GWH Model (new industry classification after using the kmeans clustering algorithm to obtain 30 groups,d=dl)

Parameter	Mean	Std.	t-statistic	Autocor.
Intercept	0.9236	33.8823	0.2552	0.1013
ME	-0.0422	1.2593	-0.3142	0.0467
E/P	-9.7024	22.5445	-4.0374	0.0679
BE/ME	-0.6736	1.3192	-4.7899	0.1189

Table 8.13: Average of Parameters estimated by the GWH Model (new industry classification after using the kmeans clustering algorithm to obtain 30 groups, d=d2), t-statistics, and Autocorrelation Values across 88 datasets

Measure of Fit	Mean	Std.
Mean-AIC	7.0057	0.4006

Table 8.14: Average of Mean AIC of GWH Model (new industry classification after using the kmeans clustering algorithm to obtain 30 groups, d=d2)

Parameter	Mean	Std.	t-statistic	Autocor.
Intercept	0.1069	33.7621	0.0297	0.0794
ME	-0.0170	1.2512	-0.1277	0.0174
E/P	-8.1943	20.8745	-3.6825	0.0644
BE/ME	-0.7804	1.3720	-5.3360	0.1440

Table 8.15: Average of Parameters estimated by the GWH Model (new industry classification after using the kmeans clustering algorithm to obtain 40 groups, d=d1), t-statistics, and Autocorrelation Values across 88 datasets

Measure of Fit	Mean	Std.
Mean-AIC	6.9975	0.3978

Table 8.16: Average of Mean AIC of GWH Model (new industry classification after using the K-means clustering algorithm to obtain 40 groups,d=d1)

Parameter	Mean	Std.	t-statistic	Autocor.
Intercept	0.4274	33.2735	0.1205	0.0872
ME	-0.0288	1.2389	-0.2182	0.0304
E/P	-7.7908	20.9622	-3.4865	0.0317
BE/ME	-0.7615	1.3212	-5.4073	0.1398

Table 8.17: Average of Parameters estimated by the GWH Model (new industry classification after using the kmeans clustering algorithm to obtain 40 groups, d=d2), t-statistics, and Autocorrelation Values across 88 datasets

Measure of Fit	Mean	Std.
Mean-AIC	6.9878	0.4065

Table 8.18: Average of Mean AIC of GWH Model (new industry classification after using the kmeans clustering algorithm to obtain 40 groups, d=d2)

The tables above show that the parameter estimates are particularly sensitive to the specification of the industry classification when the number of groups is small. When the number of groups is 10, the average of the E/P parameter is -9.8121 when we use the groups given by the k-means algorithm with distance measure d_1, but is −2.2849 when we use the groups given by the k-means algorithm when the distance measure is d_2. This even though the t-statistics for each parameter are significant (-4.05 and -4.00 respectively). Also, the average mean AIC is lower when we use d_1.When the number of groups is larger (30 or 40), the difference between the parameter estimates, appears negligeable, though the mean average AIC is lower for d_1 with 30 groups, and lower with d_2 with 40 groups. Unfortunately, no consistent pattern can recommend one distance measure over the other. At the least, we can say that when the number of groups is small (less than 20), it is better to use d_1, and when the number of groups is large, both distance measures perform about equally. Finally, it was disappointing to see that the clustering results did not allow us to obtain lower mean AIC's than the original specification. This says that the original classification is optimal, compared to all other classifications we have offered so far, whether using $d_1$or d_2, and whether making 10, 20, 30 or 40 groups out of the original 54.

8.3 Clustering Based on a Distance Defined in Terms of the Comparison of Parameters from the General Model

In this section we present the results from applying the K-means clustering algorithm to the general model of stock returns. We only work with the dataset Japan, 1993, January, and only keep sectors with more than 30 observations, as in our previous discussion of the general model. We choose our sector variables to be the parameters of the general model estimated for the original set of sectors. Each sector i in the original set of industry sectors has a unique mean term α_i and a unique variance term γ_i. These terms are standardized, as shown earlier, and the Euclidean norm of the standardized parameters is the distance between two sectors. With the K-means algorithm, we form, from the original 21 sectors, 15, 10, 5 clusters, and 1 cluster. Each time use the results of the K-means algorithm to recompute the model paramaters, and record the model AIC.

The following table gives the mapping of the original 21 groups into, respectively, 15, 10, 5 clusters, and then 1 cluster. Of course, the 1 cluster case just means that all sectors are in the same cluster.

Below, we show how the AIC changes when we use the original 21 sectors, and then 15, 10, 5, clusters, and finally just one cluster. The AIC decreases with 15, 10 and 5 clusters, with 5 clusters being the lowest. The AIC from just one cluster jumps up to 8274, which is higher than for the original 21 sectors. It therefore seems clustering on the basis of model parameters offers the best alternative for decreasing the model AIC. Unfortunately, this method is also the most computationally intensive, due to the necessity to calculate the model parameters using maximum likelihood.

Sector Number	Mapping, K=15	Mapping, K=10	Mapping, K=5	Mapping, K=1
4	5	3	1	1
8	1	5	1	1
9	3	4	3	1
12	2	2	3	1
16	6	2	3	1
17	4	4	1	1
21	7	6	2	1
24	8	1	1	1
25	9	7	4	1
28	10	8	2	1
36	3	4	3	1
41	2	2	3	1
43	11	1	1	1
44	12	9	5	1
49	13	1	1	1
54	14	10	4	1
57	1	5	1	1
59	1	5	1	1
61	15	3	1	1
63	1	5	1	1
64	4	4	1	1

Table 8.19: K-means Mapping of Original 21 Sectors: K = 15, 10, 5, 1.

AIC, Original Set of Sectors	**AIC, K=15**	**, K=10**	**, K=5**	**, K=1**
8223	8199	8179	8166	8274

Table 8.20: AIC Using K-means Results: K = 15, 10, 5, 1.

Chapter 9

Conclusion and Future Research

9.1 Summary and Conclusion

In this thesis we presented models that described the monthly cross-sectional behavior of stock returns. An underlying concern behind our modeling effort has been to obtain efficient estimates of the parameters that affect the firm-specific variables market equity (ME), earnings to price (E/P) and book to market (B/M), in models of stock return. These parameters have been shown to be significant determinants of return, based on the OLS regressions of returns on firm-specific variables.We were interested in improving on the classical OLS regressions, whose results have stimulated the writing of many papers. Our idea was to use industry classification information to model the inherent heteroskedasticity of returns across sectors. We found that on average, models that incorporated this information had lower AIC's, and therefore were better describing the mechanisms of price determination. Such was the case with OLS regression with sector dummy variables, and the groupwise hetereskedastic model (GLS), that explicitely modeled variation in variance across sectors.

Next, we tried to improve on the quality of our models, by attempting to condense the industry classification information by means of clustering algorithms. The main clustering algorithm used here has been the K-means algorithm, which requires that the final desired number of clusters be specified. Our results showed that when clustering based on the sector specific variables, there seems to be no improvement in the average AIC value of our models. Limited testing on one month of data, however, suggest that clustering based on the parameters our most general model of stock returns, which allows return variance to depend on firm-specific variables, improves the AIC.

At this point, our limited objective, to improve upon simple OLS regression of return on variables, has been achieved. Our models perform significantly better than naive applications of regression. But even though we are improving in terms of AIC, the parameter values that affect the firm-specific variables are unstable, and vary highly from month to month, although they all are significantly different from 0, based on t-statistics of their average across all months of data. Also, variation of the parameter values between models prompts us to recommend caution when trying to establish conclusions on the significance or non-significance of such variables in the determination of return.

9.2 Future Research

We forsee two main extensions to this research. The first has to do with the appropriateness of the AIC as a measure of fit of our models. Though theoretically justifiable, the AIC is just one criterion we could have used. It would be nice to see if models selected based on the AIC, and especially industry classifications chosen via the AIC, can in real life settings improve our decision making process. In particular, it would be extremely valuable to come up with an objective function that used firm-specific and industry classification information to make money, hedge risk, or that otherwise reflected some real-life financial concerns. Then, models and classifications chosen on the basis of the AIC could be checked for their empirical money making, or risk-decreasing qualities. This would in perhaps validate our statistical modeling.

The second main extension of this research would be the incorporation of the time-series behavior of stock returns. Though an orthodox view of market efficiency would say that individual stock returns evolve through time according to a random walk, recent evidence has suggested that this may not always be the case, especially for indices, such as the S&P 500[1]. One could imagine that the returns of industry groups might also be correlated through time. Also, in practice, returns may exhibit some sort of structured behavior, such as that postulated by arbitrage pricing theory (APT). In this case, there is an inherent advantage in knowing what

[1] see for example Jegadeesh (1990), and Lo and Mackinlay (1997)

that structure is. And different modeling hypotheses could be tested One possible regression framework would be to assume that the model parameters inherent to industry groups drift through time, but exhibit some degree of correlation from month to month. Such a set up can be reformulated nicely into a bayesian regression framework, and lends itself to efficient estimation using, for example, Kalman Filtering methods.

Appendix A

Industry Classifications: a Quick Overview

Industry classifications of stocks are usually based on broad economic characteristics,such as the type of activity which the firm is involved in. For example, in the government's Standard Industrial Classification - SIC -, industries are grouped by the type of products they manufacture. Others, such as the summary classification of stocks offered by Peter Lynch in "One Up on Wall Street", are based more on the industrial organization aspect of firms. In the following sections, we explain the nature of several of the most common classifications, including the Vestek one which we use in this thesis for the purpose of model implementation.

In the SIC, firms are assigned four digit codes by the Executive Office of the President Office of Management and Budget. Codes are defined in accordance with the composition and structure of the economy, and are revised periodically to reflect the economy's industrial organization. They were developed and are maintained by a commitee consisting of senior economists, statisticians, and representatives of federal agencies that use the SIC, such as the Department of Transportation.

These codes provide a standard for grouping firms. The first two digits represent a broad industrial class. They cover the entire field of economic activities, namely, agriculture, forestry, fishing, hunting, and trapping; mining; construction;, manufacturing; transportation; communication; electric, gas, and sanitary services; wholesale trade; retail trade; finance, insurance, and real estate; personnal business; professional, repair, recreation and other services; and public administration. For example, if the first two digits are 15, a firm is included in the class of building contractors. The third digit represents "industrial groups", and the fourth digit is the final industrial code assigned to each firm. Each activity listed under the four digit classification scheme must be statistically significant in the number of persons employed, the volume of the business conducted, and other measurable economic activity characteristic. Examples of these codes reveal the extent to which the division of all firms into groups is fine. A code of 152 contains single family building contractors. JC Penney and Neiman Marcus both have the code 1311, corresponding to department stores, even though their respective markets tend to be non-overlapping, Neiman Marcus being luxury goods store, and JC Penney being a standard department store. The first digit six represents insurance, and the code 6311 represents cat and dog insurance. The first two digits 02 represent agricultural production, livestock, and animal specialties, and the code 0279 represents rattle snake farms.

Standard and Poor's reports the performance for 100 groups, and calculates a stock price index for each. Value line reports on the conditions and prospects of 1700 firms, grouped into 90 industries.

Peter Lynch in "One Up on Wall Street" groups stocks according to their sensitivity to the business cycle. He defines the following five groups. Slow Growers: large, aging firms; Stalwarts: large, well-known firms; Fast Growers: small and agressive new firms, with growth rates between 20 and 25 %; Cyclicals: firms with predictable business cycles; Turnarounds: in or near bankruptcy.

Appendix B

Non-hierarchical Clustering Algorithms

We provide a brief overview of non-hierarchical clustering algorithms, and their mathematical formulations.A background knowledge of some of the vocabulary used in the Clustering literature will be useful. Consider a universe composed of points, which are usually vectors of some specified length r. A cluster is defined as a collection of such points. The distance between a point and a cluster is the distance between the point and the cluster's center - or prototype (in Pattern Recognition) . This cluster center is often the centroid of the cluster, which is a point whose coordinates are the weighted coordinates of the points within the cluster. Typically, the Euclidean distance is used, but other measures may be more appropriate in some settings. Clustering algorithms typically try to optimize some objective function that depends on the nature of the clusters in the universe of points. These algorithms sometimes, but not always, iteratively assign points to different clusters, in search of the optimal assignment of points to clusters. Clustering points together is sometimes called "clubbing'.

Issues to keep in mind when studying clustering algorithms.

1. Why is clustering being done?

2. What are the points - i.e. what are the coordinates, and do they make sense, given the following issues ?

3. What is the distance measure between points?

4. W hat is the objective function?

5. How many clusters should there be?

B.1 K-means Algorithm

One standard clustering algorithm is the K-means algorithm. The algorithm is used to solve the following mathematical program, along with its discreteness constraints:

$$\begin{array}{ll} \textit{minimize} & J(U,v) = \sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik})(d_{ik})^2 \\ s.t. & \sum_{i=1}^{c} u_{ik} = 1, \forall k = 1, ..., n \\ & u_{ik} \in \{0,1\}, \forall i = 1, ..., c, \forall k = 1, ..., n \\ & (d_{ik})^2 = (x_k - v_i)' A_i (x_k - v_i), \forall i, \forall k \end{array}$$

where U is a $(c \times n)$ matrix of weights u_{ik}, and v is $(p \times c)$ matrix whose ith colum v_i is a vector representing the cluster center of cluster i. i is the subscript indicating the cluster, c is the number of clusters, k is the subscript indicating the point, n is the number of points, u_{ik} is equal to 1 if point k belong to cluster i, and is equal to 0 otherwise. x_k, an p-dimensional vector, is the k^{th} data point, or feature vector for point k. $(d_{ik})^2$ is the distance from point k to cluster i, defined in terms of a positive definite symmetric matrix A_i. Optimality conditions imply that

$$v_i = \frac{\sum_{k=1}^{n} (u_{ik})\ x_i}{\sum_{k=1}^{n} (u_{ik})} ,$$

i.e., the v_i are cluster centroids. This last fact is the basis for the K-means algorithm, that works as follows:

Step1. Select initial location of cluster centers.

Step 2. Generate a (new) partition by assigning each point to its closest cluster center.

Step 3. Calculate new cluster centers as the centroids of the clusters.

Step 4. If the cluster partition is stable, stop; else go to Step 2.

Notice that the number of clusters to be found is fixed in advance. Each point is assigned to one and only one cluster, so both the initialization and the result of the K-means algorithm consist in a partition of the universe of points. The algorithm produces a partition which is often near optimal, yet not necessarily optimal. Solving the problem with the discreteness constraints exactly is typically very difficult given the nonlinearity of the constraints.

B.2 Fuzzy c-means Algorithm

If the discreteness constraints are relaxed, the above problem can be solved exactly using nonlinear programming theory, but the constraint set becomes intractable because of its size. Accordingly, an extension of the K-means algorithm, fuzzy c-means clustering, is used - see Bezdeck (1981). The fuzzy c-means algorithm is a heuristic to solve the following problem:

$$\begin{array}{ll} \textit{minimize} & J(U,v) = \sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^m(d_{ik})^2 \\ s.t. & \sum_{i=1}^{c} u_{ik} = 1, \forall k = 1,...,n \\ & u_{ik} \in [0,1], \forall i = 1,...,c, \forall k = 1,...,n \\ & (d_{ik})^2 = (x_k - v_i)' A_i (x_k - v_i), \forall i, \forall k \end{array} \tag{B.1}$$

The number m is a constant included in the problem definition. Different m's will give different results. The u_{ik}'s can be interpreted as probabilities, but the objective function $J(U,v)$ is not based on a probabilistic interpretation of the clustering problem. For given cluster center v_i, application of the Lagrangian multiplier theorem to the above problem yields the solution .

$$u_{ik} = \frac{1}{\sum_{j=1}^{n}[(d_{ik})^2/(d_{jk})^2]^{1/(m-1)}} \tag{B.2}$$

Also, for a given U, application of the Lagrangian multiplier theorem yields

$$v_i = \frac{\sum_{k=1}^{n}(u_{ik})^m \; x_i}{\sum_{k=1}^{n}(u_{ik})^m} \tag{B.3}$$

which says that v_i is the weighted centroid of cluster i.

The fuzzy c-means algorithm uses the two optimality conditions above, and works as follows, assuming that the numbers m and c are fixed:

1. Step 1. Select initial location of cluster centers.
2. Step 2. Generate a (new) partition using B.2
3. Step 3. Calculate the new cluster centers as the weighted centroids defined in B.3.
4. Step 4. If the cluster partition is stable, stop; else go to Step 2.

B.3 Noise in Clustering

In certain situations where noisy data is of concern, the K-means or c-means algorithms will not perform well. The basic algorithms are improved by adding a noise cluster such that the distance from any point to the noise cluster is a constant, call it δ - see Dave (1991). In effect, both algorithms remain unchanged, but the problem definition as shown above is changed so that distances become

$$\begin{aligned}(d_{ik})^2 &= (x_k - v_i)'A_i(x_k - v_i), \forall i, = 1, ..., c-1, \forall k \\ (d_{ck})^2 &= \delta^2\end{aligned} \tag{B.4}$$

Here, $c-1$ is the number of clusters, with cluster c being the noise cluster. Notice that δ must be specified in Step 1 in the above algorithm descriptions. This specification is not easy, and will depend on the problem. One scheme is to select the following statistical average,

$$\delta^2 = \lambda \left[\frac{\sum_{i=1}^{c-1} \sum_{k=1}^{n} (d_{ik})^2}{n(c-1)} \right]$$

Then, δ can be recomputed at Step 3 of the algorithm. The "noise clustering algorithm", as it is called, performs better than the K-means or fuzzy c-means on certain datasets described in Dave (1991), as measured by the location and shape of the clusters which it finds.

B.4 Possibility Theory and Clustering

Fuzzy clustering avoids having to commit a point to a single cluster. Fuzzy set methods were originally developed using membership functions. These membership functions are absolute, and denote degrees of belonging or typicality. The membership value of a point in a fuzzy set does not depend on the its value in other fuzzy sets. The fuzzy C-means algorithm, however, solves a problem where the membership values u_{ik}'s have probability interpretations because of the constraints

$$\begin{aligned} &u_{ik} \in [0,1], \forall i = 1, ..., c, \forall k = 1, ..., n \\ &\textstyle\sum_{i=1}^{c} u_{ik} = 1, \forall k = 1, ..., n \end{aligned}$$

and these values cannot be interpreted as typicality values. The mathematical program which the c-means algorithm seeks to solve can be modified to allow a possibilistic interpretation of the u_{ik} parameters - see Krishnapuram and Keller (1993). Specifically, consider the following:

$$\begin{array}{ll} \textit{minimize} & J(U,v) = \sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ik})^m (d_{ik})^2 + \sum_{i=1}^{c} \eta_i \sum_{k=1}^{n} (1 - u_{ik})^m \\ \textit{s.t.} & u_{ik} \in [0,1], \forall i = 1, ..., c, \forall k = 1, ..., n \\ & 0 < \sum_{k=1}^{n} u_{ik} \leq n, \forall i \\ & \max\limits_{i} u_{ik} > 0, \forall k \\ & (d_{ik})^2 = (x_k - v_i)' A_i (x_k - v_i), \forall i, \forall k \\ & v_i = \dfrac{\sum_{k=1}^{n} (u_{ik})^m \; x_i}{\sum_{k=1}^{n} (u_{ik})^m} \end{array}$$

where η_i are suitable positive numbers. For given centroids v_i, and relaxing the last constraint in the above problem, application of the Lagrangian multiplier theorem to the above problem yields the solution .

$$u_{ik} = \frac{1}{1 + \left(\dfrac{(d_{ik})^2}{\eta_i} \right)^{1/(m-1)}} \ . \tag{B.5}$$

The fuzzy C-means algorithm can then be appropriately modified, and the u_{ik}'s obtained can be interpreted as possibilities rather than probabilities.

Bibliography

[1] Akaike, H., 1973, "Information Theory and an Extension of the Maximum Likelihood Principle", in 2nd International Symposium on Information Theory, eds. B.N., Petroc and F. Caski, Budapest, Akedemiai Kiado, pp. 267-281.

[2] Arnott, Robert D., "Cluster Analysis and Stock Price Comovement," Financial Analysts Journal, November-December 1980, pp. 56-62.

[3] Brown, Stephen J., and William N. Goetzmann, "Mutual Fund Styles," unpublished article, available on the net at the following address: " http://pearljam.som.yale.edu/pub/wrkpaprs/goetzmann/goetzwp.htm".

[4] Bezdek, James C., 1981, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, New York.

[5] Chen, N. F., R. Roll, and S. A. Ross, "Economic Forces and the Stock Market," The Journal of Business, Vol. 59, No.3, July 1986, pp. 383-404.

[6] Connor, Gregory, "Ther Three Types of Factor Models: A Comparison of their Expllanatory Power," Financial Analysts Journal, May-June 1995.

[7] Connor, G., and R. A. Korajczyk, "A Test for the Number of Factors in an Approximate Factor Model," The Journal of Finance, Vol 48, No. 4, September 1993, pp. 1263-1292.

[8] Dave, Rajesh N., "Characterization and Detection of Noise in Clustering", Pattern Recognition Letters, Vol. 12, November 1991, pp. 657-664.

[9] Elton, Edwin J., Martin J. Gruber, and Jianping Mei, “Cost of Capital Using Arbitrage Pricing Theory: A Case Study on Nine New York Utilities,”, Financial Markets, Institutions, and Instruments, Vol. 3, No. 3, August 1994.

[10] Fama, Eugene F., and Kenneth R. French, “Common Risk Factors in the Returns on Stocks and Bonds,” Journal of Financial Economics, Vol. 33, 1993, pp. 3-56.

[11] Fama, Eugene F., and Kenneth R. French, “The Cross-section of Expected Stock Returns,” The Journal of Finance, Vol. 47, No. 2, 427-466.

[12] Farrell, James L. Jr., “Analyzing Covariation of Returns to Determine Homogeneous Stock Groupings,” Journal of Business, April 1974, pp. 186-207.

[13] Greene, William H., 1993, *Econometric Analysis*, Macmillan Publishing Company.

[14] Hartigan, John A., 1975, *Clustering Algorithms*, New York, John Wiley and Sons.

[15] Jegadeesh, N., 1990, "Evidence of Predictable Behavior of Security Returns", Journal of Finance 45, pp. 881-898.

[16] Krishnapuram, Raghu, and James M. Keller, "A Possibilistic Approach to Clustering", IEEE Transactions on Fuzzy Systems, Vol. 1, No. 2, May 1993, pp. 98-110.

[17] Lo, A., and C. MacKinlay, 1997, "Maximizing Predictability in the Stock and Bond Markets", to appear in Macroeconomic Dynamics.

[18] King, Benjamin F., “Market and Industry Factors in Stock Price Behavior,” Journal of Business, Vol. 39, January 1966, pp. 139-190.

[19] Merton, Robert C., “An Intertemporal Capital Pricing Model,” Econometrica, Vol. 41, 1973.

[20] Rosenberg, B. A., “Extra Market Components of Covariance in Security Returns,” Journal of Financial and Quantitative Analysis, Vol. 9, No. 2, March 1974, pp. 263-273.

[21] Ross, S. A., “Arbitrage Theory of Capital Asset Pricings,” Journal of Economic Theory, December 1976.

[22] Sharpe, William F., "Asset Allocation: Management Style and Performance Measurement," Journal of Portfolio Management, Winter 1992, pp. 7-19.

[23] SPlus for Windows, Version 3.1, Statistical Sciences, Inc., 1993.