

Information theoretic measure of complexity and stock market analysis :

Using the JSE as a case study

Mini Thesis in partial fulfillment of the requirements for
an M.Phil degree in Mathematics of Finance

Supervisors:

Prof. C.G.Troskie

A.E.Clark

University of Cape Town

Adeola Oyenubi

Feb 2010

PLAGIARISM DECLARATION

1. I know that plagiarism is wrong. Plagiarism is to use another's work and pretend that it is my own.
2. I have used the vancouver referencing style as the convention for citation and referencing. Each contribution to, and quotation in this mini thesis from the work(s) of other people has been contributed, and has been cited and referenced.
3. This mini thesis is my own work.
4. I have not allowed, and will not allow, anyone to copy my work.

Signature: Date:

Abstract

Bozdogan [8] [6] [7] developed a new model selection criteria called information measure of complexity (ICOMP) for model selection. In contrast to Akaike's [1] information criterion (AIC) and other AIC type criteria that are traditionally used for regression analysis, ICOMP takes into account the interdependencies of the parameter estimates.

This paper is divided into two parts. In the first part we compare and contrast ICOMP with AIC and other AIC type selection criterion for model selection in regression analysis involving stock market securities.

While in the second part we apply the definition of information theoretic measure of complexity to portfolio analysis. We compare the complexity of a portfolio of securities with its' measure of diversification (PDI) and examine the similarities and differences between the two quantities as it affects portfolio management.

Contents

I	MODEL SELECTION AND INFORMATION CRITERIA IN STOCK MARKET ANALYSIS	7
1	INTRODUCTION	7
2	VARIABLE SELECTION PROCEDURES	8
2.1	Stepwise procedures	8
2.2	Forward Selection	8
2.3	Backward Selection	9
2.4	Stepwise Selection	9
2.5	All possible Regression	9
2.6	Pros and cons of selection procedures	10
3	SELECTION CRITERIA	10
3.1	The Kullback-Leibler Distance	11
3.2	Akaike's Information Criterion (AIC)	11
3.3	Bayesian Information Criterion (BIC)	12
4	THEORY AND DEFINITION OF INFORMATION COMPLEXITY(ICOMP)	12
4.1	Definition of information complexity	12
4.2	Information complexity criteria (ICOMP)	13
4.3	Complexity and collinearity	15

5	COLLINEARITY AND MULTIPLE INDEX MODELS	16
5.1	Sharpe's Multiple Index Model	16
5.2	Detecting multicollinearity (using VIF)	17
5.3	Multicollinearity and relevant JSE economic factors	19
5.4	Effects of Multicollinearity in Multiple index Models	20
5.4.1	Methodology	20
5.4.2	Findings	21
5.5	Implications of the presence of collinearity in portfolio analysis	21
6	MODEL SELECTION, COLLINEARITY AND MULTIPLE INDEX MODELS	22
6.1	Justification of analysis	22
6.1.1	Methodology	23
6.1.2	Findings (Selected model)	23
6.1.3	Findings (Collinearity levels)	25
6.2	Discussion on models selected by AIC, BIC and ICOMP and implications in portfolio analysis	25
II	INFORMATION THEORETIC MEASURE OF COMPLEXITY AND PORTFOLIO MANAGEMENT	26
1	DIVERSIFICATION AND COMPLEXITY MEASURES FOR PORTFOLIOS	26
1.1	Measures of Diversification	26
1.2	Relationship between diversification and complexity	28

1.2.1	Definition of Covariance complexity	28
1.2.2	Maximum Covariance complexity	29
1.2.3	PDI and Complexity Measure	29
2	DIVERSIFICATION AND COMPLEXITY MEASURE OF PORT- FOLIOS USING JSEDATA	30
2.1	Methodology	31
2.2	Findings	34
2.3	Implications of findings	35
2.4	Conclusion	35
3	Appendix	40

Part I

MODEL SELECTION AND INFORMATION CRITERIA IN STOCK MARKET ANALYSIS

1 INTRODUCTION

Multiple regression analysis is very useful in many scientific fields of endeavor. Of interest to us in this mini thesis is its application to stock market data. Linear regression analysis is often used to analyze the time series of returns of securities in an attempt to explain the variation in market indices using securities.

Another typical use of regression analysis is the Sharpe's simple and multiple index models (covered in section 5), where we are interested in the market factors (e.g. indices, interest rates etc.) that can be used to explain the variation in security returns.

In a multiple regression type problem it is often useful to be able to trim down the number of variables that are needed to explain the variation in a response variable. To achieve this some form of information criteria is used as a goodness of fit measure to determine the best subset of available explanatory variables that can explain the variation in the response variable. Occum's Razor [4] suggested that "simpler" models should be preferred to complex ones. Complexity as suggested by Occum's Razor is in terms of the number of variables included in the model.

According to Bozdogan [10] complexity is the measure of the degree of interdependency between the whole system and a simple enumerate composition of its subsystem. This definition has been explained to suggest that the complexity of a model is not solely a function of dimension but also how related the variables are to each other (interdependency).

In this mini thesis we compare the performance of AIC type model selection criteria (constant penalty term) with the one based on information complexity (dynamic penalty term based on dependency) and further discuss the advantages and disadvantages of the criteria in the presence of collinearity as it applies to stock market data analysis. In order to do this we establish the presence of collinearity in stock market data.

We first discuss model selection criteria and selection procedures. Then we present what information complexity is. The results of various analyses done on JSE data using the various information criteria is then presented.

2 VARIABLE SELECTION PROCEDURES

The first thing to do in a model selection problem is to decide on the selection criteria to be used. A selection criterion is a rule that is use to decide a suitable model among a set of competing models (Details on selection criteria is presented in the next section). After this step a decision is made on the selection procedure to be employed in determining the best subset of variables that explains variation in the response variable.

2.1 Stepwise procedures

There are three types of stepwise procedures

1. Forward Selection
2. Backward Selection
3. Stepwise Selection

2.2 Forward Selection

In the forward selection procedure the selection criteria is computed for all the available explanatory variables. The variable that produces the best selection criteria

value is selected. This process is repeated and another variable is added if it improves the value of the selection criteria given that the first variable selected is still in the model. This process is repeated until we can no longer find a variable that result in an improved selection criteria value in the presence of all other variables previously selected.

2.3 Backward Selection

We start this procedure by including all explanatory variables in the model. Then each of these variables are considered as if it is the last variable to enter the model. The variable which when removed result in the greatest improvement in the selection criteria value is removed from the model. The process is repeated until we can no longer find a variable that result in an improvement on the selection criteria value when removed from the model.

2.4 Stepwise Selection

This is a combination of the forward and backward procedure. It starts off by selecting the variable that result in the best selection criteria value. However at each step after the first step the procedure looks backwards removing the variable that was earlier selected but when removed results in an improved selection criteria value. The process is terminated when we can no longer find a variable that results in an improvement in the selection criteria value by removing variables earlier selected or by adding variables that has not been selected in the model.

2.5 All possible Regression

As the name suggests we simply fit all subsets including the full set and select the model that gives the best selection criterion value.

2.6 Pros and cons of selection procedures

Forward, Backward and Stepwise procedures in regression analysis do not always find the best subset of predictor variables from the set explanatory variables. Some of the problems with these procedures include:

1. Little or no theoretical justification exists for the order in which variables enter or exit the algorithm. (Boyce et al. [5], L. Wilkinson [29]).
2. Stepwise searching rarely finds the overall best model or even the best subset of a particular size. (Mantel [16], Hocking [25]).
3. Stepwise selection provides extremely limited sampling from a small area of the vast solution space, at the very best, it can only produce an “adequate” model. (Sokal and Rohlf [24]).

All possible regressions on the other hand has a disadvantage of being time consuming. All subset regressions is the preferred procedure because it examines all possible combinations of variable before making a decision. When this procedure cannot be used, stepwise selection procedure is used as it has the advantage of sampling from a larger solution space than forward and backward selection.

3 SELECTION CRITERIA

According to Linhart and Zucchini [15], model selection can be formally separated into two sages. The first stage involves a decision on the choice of approximating family e.g multiple regression. The second step entails estimating the parameters of the model in order to compare competing models. Kullback and Leibler [14] states that any selection criteria should ensure that that the final model selected should be as close as possible to the true data generating process of the data.

3.1 The Kullback-Leibler Distance

Let $f(x)$ be the density function of the true data generation process we wish to model where \mathbf{X} is a multivariate random variable. In reality $f(x)$ is unknown and has to be approximated by using a finite set of data. Let $g_i(x)$ (for $i=1 \cdots m$) be a set of competing models that approximate $f(x)$. In an attempt to measure the suitability of an approximating model the Kullback-Leibler (KL) distance was derived Kullback and Leibler [14]. The KL distance measures how well $g_i(x)$ approximates $f(x)$.

Based on the above principle, various selection criteria were developed. It is a value that allows one to rank models in terms of how well the explanatory variables perform in approximating the dependent variable. This can be easily determined by using informational model selection criteria. For the purpose of this mini thesis we discuss three informational model selection criteria

1. Akaike information criterion, AIC
2. Schwarz's information criterion, BIC
3. Information Complexity criterion, ICOMP

Many model selection criterion take the form of a penalized likelihood.

3.2 Akaike's Information Criterion (AIC)

Akaike [1] showed that the KL distance can be approximated by

$$AIC = -2\log L(\theta) + 2k \quad (1)$$

where $L(\theta)$ is the likelihood function evaluated at θ , θ is the maximum likelihood estimate (MLE) of the k parameters of the model.

The model that minimizes AIC is preferred since it would theoretically be the closest model to the unknown model. AIC is composed of two parts the namely, "the goodness of fit" (the maximum likelihood) of the model and the number of free parameters in the model. The second term in AIC acts as a penalty function which penalises high

dimensional models more heavily than low dimensional models. Here the dimension of a model is in terms of number of parameters in the model.

3.3 Bayesian Information Criterion (BIC)

This is often referred to as a dimension consistent information criterion i.e as the number of observations becomes large, the correct model is selected with probability one. Schwarz [22] developed the Bayesian Information Criterion (BIC). The criterion is defined as

$$BIC = -2\log L(\theta) + k\log(n) \quad (2)$$

Where n is the number of observations in the data set. Model selection is undertaken by selecting the model that minimises BIC.

4 THEORY AND DEFINITION OF INFORMATION COMPLEXITY(ICOMP)

4.1 Definition of information complexity

One of the desirable properties of model selection criteria is to ensure that “simpler” models are preferred to more complex ones. This allows for better interpretation of the model. By simpler models the traditional definition refers to the dimension of the model (parsimony).

Another definition of complexity is that “It is the measure of the degree of interdependency between the whole system and a simple enumerate composition of its subsystem” [10]. This definition suggests that complexity of a model is not solely a function of dimension but also how related the variables are to each other (interdependency).

According to the same authors “The general principle is that for a given level of accuracy, a simpler model (i.e one with a small covariance matrix of the parameter

estimates and residual covariance matrix) is preferred to more complex one. Here small is used in the sense of minimum variance.”

4.2 Information complexity criteria (ICOMP)

Bozdogan has used the above definition of complexity to derive a number of Information Complexity Criteria (ICOMP) used for model selection. ICOMP uses a penalty function that is based on the information complexity index of Van Emden [28]. ICOMP penalizes the covariance complexity of a estimated model.

Models that have many insignificant parameter estimates as well as collinear data sets are penalized more heavily than models containing many significant parameter estimates based on an orthogonal data matrix.

The procedure ICOMP as presented by Bozdogan [11] is based on the structural complexity of a set of random vectors via a generalization of the information-based covariance complexity index of van Emden [28].

For a general multivariate linear or nonlinear model defined by

$$\textit{Statistical model} = \textit{Signal} + \textit{Noise}$$

ICOMP is designed to estimate a loss function [11] (difference between the estimated model and the true model)

$$\textit{Loss} = \textit{Lack of Fit} + \textit{Lack of Parsimony} + \textit{Presence of Complexity}$$

$ICOMP_0$ is defined as

$$ICOMP_0 = -2\log L(\hat{\theta}) + 2C_0(\Sigma_{\hat{\theta}}) \quad (3)$$

$$C_0(\Sigma_{\hat{\theta}}) = \frac{1}{2}\log(\text{tr}(\Sigma)) - \frac{1}{2}\log(|\Sigma|) \quad (4)$$

where $\hat{\theta}$ is the estimated parameters of the model and the estimated residual variance and co-variances, k is the number of estimated free parameters in the model and $\Sigma_{\hat{\theta}}$ is the estimated covariance matrix of the model parameters and residuals. $ICOMP_0$ is not invariant under orthogonal transformations [28] but it is invariant with respect to scalar multiplication and orthogonal transformation. Bozdogan [6] [8] [9] shows

that the maximal informational complexity of $\Sigma_{\hat{\theta}}$, $ICOMP(C_1(\Sigma_{\hat{\theta}}))$, solves the above problem of $ICOMP_0$. $ICOMP(C_1(\Sigma_{\hat{\theta}}))$ is defined as

$$ICOMP(C_1(\Sigma_{\hat{\theta}})) = -2\log L(\hat{\theta}) + 2C_1(\Sigma_{\hat{\theta}}) \quad (5)$$

$$C_1(\Sigma_{\hat{\theta}}) = \max_T C_0(\Sigma_{\hat{\theta}}) \quad (6)$$

$$= \frac{p}{2} \log \left(\frac{\text{tr}(\Sigma_{\hat{\theta}})}{p} \right) - \frac{1}{2} \log(|\Sigma_{\hat{\theta}}|) \quad (7)$$

where the maximum (in equation 6) is taken over orthonormal transformation T of the overall coordinate systems of the parameters and p is the dimension of $\Sigma_{\hat{\theta}}$.

$ICOMP(IFIM)$ is defined as

$$ICOMP(IFIM) = -2\log L(\hat{\theta}) + 2C_1(F^{-1}(\theta_k)) \quad (8)$$

Where $C_1(F^{-1})$ is the maximal information complexity of F^{-1} , the inverse of the estimated Fisher Information matrix. F^{-1} gives a scalar measure of the Cramer-Rao lower bound matrix which takes into account the accuracy of the estimated parameters and implicitly adjusts for the number of free parameters included in the model (Cramer [13]) and (Rao [18] [20] [19]).

The use of F^{-1} in the information-theoretic model evaluation criteria takes into account the fact that as we increase the number of free parameters in a model, the accuracy of the parameter estimates decreases.

The measure based on F^{-1} is preferred based on the principle of parsimony, $ICOMP(IFIM)$ chooses simpler models that provide more accurate and efficient parameter estimates over more complex, over specified models.

We note that the trace of IFIM in the complexity measure involves only the diagonal elements; analogous to variances, while the determinant involves the off diagonal elements; analogous to covariances. Therefore, $ICOMP(IFIM)$ contrasts the trace and the determinant of IFIM, and this amounts to a comparison of the geometric and arithmetic means of the eigenvalues of IFIM given by

$$ICOMP(IFIM) = -2\log L(\hat{\theta}) + 2s \log \left(\frac{\lambda_a}{\lambda_g} \right) \quad (9)$$

$$\text{where } s = \dim(F^{-1}(\theta_k)) = \text{rank}(F^{-1}(\theta_k))$$

λ_a is the arithmetic mean of eigenvalues values of F^{-1} and λ_g is the geometric mean of eigenvalues of F^{-1} . The first component of ICOMP (IFIM) is a measure of the fit of the model where as the second component measures the complexity of the estimated inverse of the Fisher Information matrix.

The greatest simplicity, that is zero complexity, is achieved when F^{-1} is proportional to the identity matrix, implying that the parameters are orthogonal and can be estimated with equal precision. In this sense, parameter orthogonality, several forms of parameter redundancy, and parameter stability are all taken into account [11].

4.3 Complexity and collinearity

In a multiple regression model collinearity refers to a situation where the design matrix is ill-conditioned. This is a situation where two or more of the explanatory variables are linearly related or very nearly linearly related. Silvey [23].

This poses a problem because as the explanatory variables become more highly correlated, it becomes more and more difficult to determine which of the explanatory variables is actually accounting for the variation in response variable.

According to Thiart [26], “collinearity cannot be described in simple terms as being present or absent. Rather, what is important is the degree and what effect this degree can have on the regression model.”

In a paper by Clark and Troskie [27] the authors through a simulation study submitted that “as the collinearity levels in the design matrix increased, the agreement percentages for all of the information criteria decreased monotonically”. Agreement percentage refers to the number of times the selection criterion under study (AIC, BIC and ICOMP) agree with the KL decision or the number of times the selected model using these criteria approximates the true data generating process (i.e chose the correct model). This means that collinearity directly affect the performance of information criterion.

It is not difficult to see that complexity and multicollinearity in linear regression models move in the same direction. For instance, in a regression model where the design matrix have highly correlated explanatory variables, we expect the covariance

and correlation matrix of the regressors to have large covariances and significant correlation values respectively. By extension we expect the covariance matrix of the parameter estimates to have large values.

5 COLLINEARITY AND MULTIPLE INDEX MODELS

In this section we highlight the use of model selection technique in stock market data analysis and discuss the presence and the effect of collinearity in this type of analysis. We then compare the performance of AIC and BIC with that of ICOMP in selecting variables in the context of multiple index models.

5.1 Sharpe's Multiple Index Model

In the context of stock market data analysis a multiple index model is simply a multiple regression model that attempts to capture some of the non-market influences that causes stocks to move together.

Covariances between security returns (in the same market or economy) tend to be positive because the same economic forces affect the fortunes of many firms. The approach tries to use indices that are not too correlated with the market index to capture additional information relevant to shares that were not contained in the market index. In contrast if we summarize all relevant economic factors by one indicator and assume that it moves all the shares in the market as a whole we will have a single Index Model.

Sharpe's Multiple Index Model can be written as [3]

$$R_{it} = \alpha_i + \beta_1 I_{i1} + \beta_2 I_{i2} + \cdots + \beta_m I_{im} + e_{it}, \quad i = 1 \cdots, m, t = 1 \cdots, n \quad (10)$$

Where R_{it} is the return of stock i in period t , α_i the unique expected return of security i , β_{i1} the sensitivity of stock i to market variable I , I_{it} the return on the market variable in period t , and e_{it} is the unique risky return of security i in period t . with the following assumptions

$$E(e_{it}) = 0, \quad \forall i$$

$$E(e_{it}^2) = \sigma^2, \quad \forall i$$

$$E(e_{it}e_{is}) = 0, \quad t \neq s, \quad s = 1, \dots, n$$

$$E(e_{it}I_{jt}) = 0, \quad j = 1, \dots, m, \quad t = 1, \dots, n$$

$$E(e_{it}e_{jt}) = 0, \quad t = 1, \dots, n, \quad i \neq j$$

$$\text{Cov}(I_{jt}I_{kt}) = c_{jk}, \quad j, k = 1, \dots, m$$

One of the major uses of the multi-Index model is to supply inputs for the Markowitz portfolio selection procedure. Markowitz's procedure involves constructing an efficient portfolio by minimizing the portfolio's risk for every level of expected returns. The number of estimates required for the Markowitz procedure using the index model is only a fraction of what otherwise will be needed [3].

To use Sharpe's multiple index model in a Markowitz procedure it is often necessary to trim down the number of indices to be used to predict the return, variance and co-variances of securities. This is because we are interested in only the important variables (factors) that can explain the variation in the stock movement. This is a typical model selection problem.

We often have a wide range of indices (economic factors) to choose from in order to model security returns. In most cases we cannot rule out dependences (collinearity) between these economic variables. Traditionally AIC and other AIC type selection criteria is used to trim down the economic variables.

In this mini thesis we compare the performance of AIC and BIC with that of selection criteria that penalizes the model based on the complexity of the parameter estimates and residual covariance matrix. First we demonstrate the presence of collinearity in this type of problem and discuss the implications.

5.2 Detecting multicollinearity (using VIF)

As stated earlier, multicollinearity cannot be described in simple terms as being present or absent the important thing is the degree and the effect of this degree

has on the regression model. There is no test that can state for sure if it is a problem or not. There are however many warning signals. The Variance Inflation Factor (VIF) is one measure that can be used to quantify the degree of collinearity in a data set. Let H = the set of all the X (independent) variables including the intercept in a linear regression model. Let G_k = the set of all the X variables except X_k . The formula for standard errors of the coefficient β_k is then

$$\text{var}(\hat{\beta}_k) = \frac{\sigma^2}{(n-1)\text{var}(X_k)} \cdot \frac{1}{1-R_k^2} \quad (11)$$

where R_k^2 is the multiple R^2 for the regression of X_k on the other explanatory variables. This means that the bigger the value of R_k^2 (i.e. the more highly correlated X_k is with the other explanatory variables) the larger the standard error will be.

Also, $1 - R_k^2$ is referred to as the tolerance of X_k . A tolerance close to 1 suggests that there is little multicollinearity in the design matrix, whereas a value close to 0 suggests that multicollinearity may be present.

The reciprocal of the tolerance is known as the Variance Inflation Factor (VIF). The VIF shows us how much the variance of the coefficient estimate is being inflated due to multicollinearity.

The square root of the VIF tells us how much larger the standard error is, compared with what it would be if that variable were uncorrelated with the other X variables in the regression equation. There are other ways of detecting multicollinearity which include

1. None of the t-ratios for the individual coefficients is statistically significant, yet the overall F statistic is.
2. Instability of the regression coefficients for different samples and different variables in the regression equation.

For the purpose of this thesis however we will make use of VIF (There are other methods suggested by Belsley, Kuh, and Welsch [2], like Singular value decomposition, Conditional index and Variance decomposition proportions). As a general rule, when VIF is greater than 10 multicollinearity is considered to be a problem (Belsley, Kuh,

Welsch, [2]).

5.3 Multicollinearity and relevant JSE economic factors

To establish the presence of multicollinearity in a multiple index problem using the JSE as a case study, we select ten economic indicators (factors or indices) that can help model the returns (calculated using equation 12) of top 50 stocks on the Johannesburg Stock Exchange (JSE).

$$r_i = \log \left(\frac{P_i}{P_{i-1}} \right) \quad (12)$$

where P_i is the price of the stock on day i . We make use of monthly data between January 2001 and July 2009. The selected variables are¹

- J203 : All Share Index
- J200 : Top 40 shares on the JSE (Tradable)
- J835 : JSE-Banks (Index containing bank shares)
- J211: JSE-INDI (Index containing Industrial shares)
- J210: JSE-RES (Index containing resources shares)
- GOLD: Gold Price (Gold price in Rand)
- JSEEX: ZAR/USD (Rand dollar exchange rates)
- R153: Yeild on rand denominated South African government bond maturing 31 August 2011
- DJT: DJ-Trans (Dow Jones transport index)
- SP500: Standard and Poor 500 index (USA)

¹This data was provided by Prof Troskie (Statistics Department UCT) and Terry Steward (Mathematics Finance Student UCT).

5.4 Effects of Multicollinearity in Multiple index Models

5.4.1 Methodology

The log returns of the variables were used in the analysis. First we compute the variance covariance matrix of the variables as presented below. The problem with using correlation matrix alone is that it can only reveal pairwise correlations. We went further to calculate the VIF of each variable by regressing it against other explanatory variables. For example to get the VIF for J203 (JSE All Share Index) we estimate the regression equation

$$J203 = \alpha + \beta_{J200}X_{J200} + \beta_{J835}X_{J835} + \beta_{J211}X_{J211} + \beta_{J210}X_{J210} + \beta_{GOLD}X_{GOLD} \\ + \beta_{JSEEX}X_{JSEEX} + \beta_{R135}X_{R135} + \beta_{DJT}X_{DJT} + \beta_{SP500}X_{SP500} + e$$

The R^2 value of this regression equation is used in the formulae

$$VIF_{J203} = \frac{1}{1 - R_{J203}^2} \quad (13)$$

Table 1 shows the correlation matrix of the factors.

Table 1:CORRELATION MATRIX FOR FACTORS

Correlation Matrix of Factors										
	J203	J200	J835	J211	J210	GOLD	JSEEX	R135	DJT	SP500
J203	1.0000									
J200	0.9971	1.0000								
J835	0.5147	0.4775	1.0000							
J211	0.8465	0.8244	0.6030	1.0000						
J210	0.9006	0.9213	0.1985	0.5648	1.0000					
GOLD	0.2582	0.2520	0.1811	0.0768	0.3041	1.0000				
JSEEX	-0.0201	0.0132	-0.3845	-0.1691	0.1696	-0.2145	1.0000			
R135	0.1285	0.1667	-0.3124	0.0409	0.2571	-0.1951	0.3218	1.0000		
DJT	0.5342	0.5276	0.3540	0.5785	0.3811	-0.0550	-0.2339	0.1058	1.0000	
SP500	0.6412	0.6228	0.4786	0.7235	0.4162	0.0350	-0.3618	0.0088	0.7541	1.0000

Table 2: VIF value for each factor computed using the equation 11

VIF (factors)	R .sq	VIF	sqr(VIF)
J203	0.9974	384.6153	19.6116
J200	0.9979	476.1904	21.8217
J835	0.0303	1.0313	1.0155
J211	0.9629	26.9541	5.1917
J210	0.9895	95.2380	9.7590
GOLD	0.3129	1.4553	1.2063
JSEEX	0.4235	1.7346	1.3170
R 153	0.3658	1.5767	1.2557
DJT	0.5938	2.4618	1.5690
SP500	0.7503	4.0048	2.0012

5.4.2 Findings

With the correlation matrix we note that some of the variable are highly correlated, for example the correlation between some of the market indices are in the order of 0.9.

The VIF figures show the levels of collinearity in the factors, we can infer from the VIF figures that four (namely J203, J200, J211 and J210) out of the ten variable have VIF values greater than 10.

In the last column we have the square root of the VIF's. The square root of the variance inflation factor tells us how much larger the standard error is, compared with what it would be if that variable were uncorrelated with the other independent variables in the regression equation. The variables written with bold font (in table 2) have multicollinearity problems and can therefore inflate the standard errors in the regression and more importantly give unstable or inaccurate values of beta.

5.5 Implications of the presence of collinearity in portfolio analysis

In a case of severe multicollinearity the issue of unstable or incorrect parameter estimates is very important. This is because in a situation where we are interested in using Sharpe's Multiple Index Model to get input estimates for a Markowitz portfolio selection problem, unstable parameter estimates implies that the estimated mean, variance and covariances obtained from the selected model can be erroneous.

The value of beta for each economic factor tells us the sensitivity of the security returns to changes in this factor. In the presence of multicollinearity these beta estimates may be highly inaccurate.

There are cases where we are interested in constructing a factor portfolio. This is a well diversified portfolio constructed to have a beta of one on one of the factors and a beta of zero on any other factor. With severe multicollinearity it is difficult to construct such portfolios.

6 MODEL SELECTION, COLLINEARITY AND MULTIPLE INDEX MODELS

6.1 Justification of analysis

To get input estimates for a Markowitz portfolio selection procedure using multiple index models we need to select important variables for the model that will be used to estimate the mean, variance and covariance of security returns.

In the last section we established the presence and the effect of multicollinearity in stock market regression analysis, specifically for Sharpe's Multiple Index Models. In this section we go on to highlight the importance of using a model selection criteria that penalizes interdependences between factors in a model. When this type of selection criteria is used as argued earlier it will select models in such a way that the design matrix is as orthogonal as possible.

The next analysis compares the performance of AIC, BIC and ICOM(IFIM) in selecting important variables for modeling security returns vis-a-vis the result of the multicollinearity levels reported in the last analysis.

In addition to data on economic factors used in the last analysis we collected monthly data on the top 50 Stocks on the JSE.

6.1.1 Methodology

To carry out the analysis we included the R package “icomp” that computes $ICOMP_1$ using complexity of the estimated inverse of the Fisher Information matrix as the penalty term. We also included the R package “MuMin” that has the function “drega” which selects variables for a given criteria using all possible regression procedure.

The log returns of the securities are used as the dependent variable in each case. The log returns of the factors are used as the explanatory variables. The *drega* routine in R is used to select the set of variables that minimizes *AIC*, *BIC* and $ICOMP_1$ using the all possible regression procedure.

Table 3 below displays the model selected when using all possible regressions for *AIC*, *BIC* and $ICOMP_1$ for some of the shares. A detailed table that shows selected models using Forward, Backward and Stepwise procedure for AIC and BIC is included in the appendix.

Table 3:REPORT AIC, BIC AND ICOMP VALUES FOR EACH OF THE
SELECTED MODELS

ALL POSSIBLE REGRESSION						
STOCK	R^2	AIC MODEL	R^2	BIC MODEL	R^2	ICOMP MODEL
ABL	0.45	DJT + J200 + J203 + J835	0.43	J835	0.46	DJT + GOLD + J203 + J835 + JSEEX + R153
ACL	0.17	GOLD + J835 + R153	0.17	GOLD + J835 + R153	0.17	GOLD + J835 + JSEEX + R153
AEG	0.40	DJT + J200 + J203 + JSEEX + R153	0.38	J200 + J203 + R153	0.27	DJT + J211 + JSEEX + R153
AGL	0.85	GOLD + J210 + SP500	0.85	GOLD + J210 + SP500	0.85	GOLD + J210 + SP500
AMS	0.71	DJT + GOLD + J210 + SP500	0.71	GOLD + J210 + SP500	0.71	DJT + GOLD + J210 + J835 + JSEEX + SP500
ASA	0.71	GOLD + J200 + J203 + J210	0.68	J835	0.69	GOLD + J211 + J835 + JSEEX + R153
ARI	0.46	DJT + J210 + JSEEX + SP500	0.44	J210 + JSEEX	0.45	DJT + J210 + JSEEX
ANG	0.51	GOLD+J203+J210+J211+J835+SP500	0.48	GOLD + J210 + SP500	0.48	GOLD + J210 + SP500
APN	0.19	J200 + J203 + J835	0.16	J835	0.18	GOLD + J835 + JSEEX + R153 + SP500
BIL	0.72	DJT + J200 + J203 + J835	0.79	GOLD + J200 + J211	0.79	DJT + GOLD + J210 + J211 + J835
CFR	0.67	J200 + SP500	0.67	J200 + SP500	0.67	GOLD + J203 + R153 + SP500
DDT	0.54	GOLD + J211 + JSEEX + R 153 + SP500	0.51	J211 + JSEEX	0.54	GOLD + J211 + J835 + JSEEX + R153 + SP500
DSY	0.31	J203 + J210 + J835 + JSEEX + SP500	0.29	J203 + J210 + J835 + SP500	0.26	J210 + J211 + J835 + JSEEX + R153 + SP500
FSR	0.89	J200 + J210 + J211 + J835	0.89	J835	0.89	J835
GFI	0.55	GOLD + J210 + JSEEX + SP500	0.54	GOLD + J210 + SP500	0.55	GOLD + J210 + JSEEX + SP500
GRT	0.43	J200 + J203 + J835 + R 153	0.43	J200 + J203 + R153	0.41	GOLD + J203 + J835 + JSEEX + R153
HAR	0.53	GOLD + J210 + JSEEX + SP500	0.51	GOLD + J210 + SP500	0.52	GOLD + J210 + JSEEX + R153 + SP500
IMP	0.65	GOLD + J200 + J211 + SP500	0.67	GOLD + J200 + J211	0.67	GOLD + J200 + J211 + JSEEX + R153 + SP500
IPL	0.45	GOLD + J200 + J203 + JSEEX + R 153	0.40	J200 + J203	0.42	GOLD + J203 + J210 + JSEEX + R153 + SP500
INP	0.51	DJT + J200 + J210 + R 153 + SP500	0.49	J200 + J210	0.501	DJT + GOLD + J200 + J210 + R153 + SP500

6.1.2 Findings (Selected model)

Refer to table 3

- ICOMP avoids selection of variables with high correlation values in the same model. For instance, ICOMP never selects J200 and J203 in the same model. These factors yield the highest pair wise correlation values and VIF figures.
- In most cases the R^2 for the selected model using ICOMP and the one using AIC or BIC are more or less the same. This implies that ICOMP select models that explain as much variation in the dependent variable as AIC and BIC while minimising the Covariance complexity of the estimated parameters.
- BIC tend to select the model with the minimum complexity (dimension) in terms of the number of factors selected while ICOMP tend to select the model with the maximum complexity (dimension). On the other hand ICOMP selects the model with the minimum complexity in terms of interdependences of parameter estimates out of the three selection criterion.

Table 4: ANALYSIS ON SELECTED MODELS

AMS											
AIC				BIC				ICOMP			
Model	R .sq	VIF	sqr(VIF)	Model	R .sq	VIF	sqr(VIF)	Model	R .sq	VIF	sqr(VIF)
DJT	0.586	2.418	1.555	GOLD	0.102	1.114	1.056	DJT	0.587	2.423	1.557
GOLD	0.128	1.147	1.071	J210	0.257	1.346	1.16	GOLD	0.266	1.362	1.167
J210	0.277	1.382	1.176	SP500	0.182	1.223	1.106	J210	0.447	1.808	1.345
SP500	0.589	2.433	1.559					SP500	0.671	3.038	1.743
								J835	0.296	1.421	1.192
								JSEEX	0.398	1.660	1.288
R^2 for the model			0.7108	R^2 for the model			0.7055	R^2 for the model			0.7101

BVT											
AIC				BIC				ICOMP			
Model	R^2	VIF	sqr(VIF)	Model	R^2	VIF	sqr(VIF)	Model	R^2	VIF	sqr(VIF)
GOLD	0.161	1.193	1.092	GOLD	0.161	1.193	1.0923	GOLD	0.289	1.407	1.186
J200	0.997	357.140	18.898	J200	0.997	357.143	18.898	J203	0.874	7.943	2.818
J203	0.996	285.714	16.903	J203	0.996	285.714	16.903	J210	0.874	7.955	2.821
J210	0.913	11.520	3.394	J210	0.913	11.520	3.394	JSEEX	0.337	1.508	1.228
								DJT	0.415	1.710	1.308
								R 153	0.236	1.308	1.144
R^2 for the model			0.629	R^2 for the model			0.629	R^2 for the model			0.604

“ R^2 for the model” gives the Coefficient of determination for the selected model under the criterion while “ R^2 ” gives the Coefficient of determination for the variable when it is regressed against other explanatory variables.

To highlight the difference in levels of collinearity in models selected by AIC and BIC compared to the one selected by ICOMP(IFIM) we selected two cases using AMS

(ANGLO PLATINUM LTD) and BVT (THE BIDVEST GROUP LIMITED).

6.1.3 Findings (Collinearity levels)

Refer to table 4 above

- As stated earlier the amount of variation explained by the three selected models are more or less the same. However ICOMP selected a more complex model in terms of number of variables than AIC and BIC for both cases.
- In the first case (AMS), there is no worrisome level of collinearity ($VIF < 10$) in the model selected by the three criteria. However in the second case (BVT), the level of collinearity in the design matrix selected by AIC and BIC is high ($VIF > 10$) compared to the one selected by ICOMP ($VIF < 10$).

6.2 Discussion on models selected by AIC, BIC and ICOMP and implications in portfolio analysis

From the above findings it is clear that the models selected by AIC and BIC tend to have higher level of collinearity in the design matrix than the one selected by ICOMP. It is also safe to say that the model selected by ICOMP has more orthogonal design matrix than the one selected by AIC and BIC. This means that the parameter estimates for model selected by ICOMP is more stable than the one selected by AIC and BIC due to the possibility of high level of collinearity in the model selected by the later especially when AIC and BIC do not agree with ICOMP. The submissions above are only valid for the data collected because it is difficult to generalize this result without conducting a simulation study. This could be considered for future work.

Stable parameter estimates implies the estimated response of the security to changes in the factors when the effect of other factors is taken into consideration is more accurate.

Part II

INFORMATION THEORETIC MEASURE OF COMPLEXITY AND PORTFOLIO MANAGEMENT

1 DIVERSIFICATION AND COMPLEXITY MEASURES FOR PORTFOLIOS

In this section we apply the definition of information complexity to portfolio risk management. According to Markowitz [17] the two important parameters to be considered by investor is the returns and the volatility (risk) of the portfolio. One of the ways of reducing the volatility (risk) of portfolio returns is to ensure that the portfolio is well diversified. Diversification involves making investments in a wide variety of assets in an effort to minimize risk, it is one of the core objectives for combining assets in a portfolio. Having investments in securities with unrelated sources of variation reduces the chances of the constituents of the portfolio moving in the same direction (either loss or gain) at any given point in time thereby reducing the over all risk of the portfolio.

Basically we investigate the effect differences in levels of interdependences among securities that makes up a portfolio has on its diversification and consequently its risk.

1.1 Measures of Diversification

Traditionally the extent to which a portfolio is diversified is measured relative to the market index. This involves regressing portfolio returns against the returns of the market index and finding the variance of the portfolio returns. The total risk of a portfolio can be partitioned into Systematic risk (or market risk) and unsystematic

risk (or Firm specific risk), i.e

$$TOTAL\ RISK = SYSTEMATIC\ RISK + FIRMSPECIFIC\ RISK \quad (14)$$

where firm specific risk is also referred to as diversifiable risk. The diversifiable risk is then taken as a measure of diversification of the portfolio because it represents the component of the portfolio's risk that can be eliminated through diversification. Thus the smaller unsystematic risk component the more diversified the portfolio [3].

It has been argued that this traditional method has a serious flaw in that it measures diversification relative to market index. What happens if the market index itself is not appropriately diversified? In smaller markets where concentration² is potentially high, measuring diversification relative to the market index will give biased results. What is needed to measure diversification concept cleanly is a measure which is independent of the overall market index. A well diversified portfolio is one in which the return and subsequently the risk arises from as many independent (unrelated) sources as possible (D. Bradfield et al. [12]).

Alexander Rudin and Jonathan Morgan [21] proposed a measure of diversification that is based on the number of independent factors observed in a portfolio. These factors are quantified using Principal Component Analysis (PCA) and they are called Portfolio Diversification Index (PDI). The procedure to be followed to calculate PDI are

- The required data is the current composition on stock level and a return history of stocks held in the portfolio.
- The time series of returns is then multiplied multiplied by their respective weights for each of the N stocks
- PCA is then conducted on the covariance matrix of the series to quantify all uncorrelated sources of risk and their relative magnitudes.

²Concentration of a portfolio refers to the extent to which portfolio weights moves away from an equally weighted distribution of securities in the portfolio.

- The factors are ordered in terms of magnitude of (using their eigenvalues) from the highest to the lowest.

This is then substituted into the following formula

$$PDI = 2 \sum_{k=1}^N k \lambda_k - 1 \quad (15)$$

Where N is the number of assets, λ_k is the percentage contribution of stock k to total volatility.

PDI is the balancing point of independent factors thus [12]

- For a completely undiversified portfolio which is dominated by a single factor the PDI is 1
- For a completely diversified portfolio the PDI is N
- The smaller the PDI measure the less diversified the portfolio.

The main idea is that a well diversified portfolio should have many equal sources of volatility emanating from several uncorrelated factors [12].

1.2 Relationship between diversification and complexity

1.2.1 Definition of Covariance complexity

Informational complexity of a covariance matrix Σ for the multivariate normal distribution with random variables x_1, \dots, x_p is defined in equation (6) (Note however that here the formulae is being applied on the covariance matrix of the returns and not the covariance matrix of estimated linear model parameters). where σ_{jj} is the j^{th} diagonal element of Σ and p is the dimension of Σ . Note that $C_0(\Sigma) = 0$ when Σ is a diagonal matrix. (i.e., if the variables are linearly independent). $C_0(\Sigma)$ is infinite if any one of the variables may be expressed as a linear function of the others ($|\Sigma| = 0$). If $\theta = \theta_1, \dots, \theta_p$ is a normal random vector with covariance matrix equal to $\Sigma(\theta)$ then $C_0(\Sigma(\theta))$ is simply the KL distance between the multivariate normal density of θ and the product of the marginal densities of the components of θ .

1.2.2 Maximum Covariance complexity

A maximal information theoretic measure of complexity of a covariance matrix Σ of a multivariate normal distribution is as defined in equations (6) and (7). $C_1(\Sigma)$ is an upper bound to $C_0(\Sigma)$ and it measures both the inequality among the variances and the contribution of the covariances in Σ (van Emden [28]). This measure of complexity provides a numerical measure to assess redundancy and stability uniquely all in one measure. We can re-write equation (7) as

$$C_1(\Sigma) = \frac{p}{2} \log \left(\frac{\lambda_a}{\lambda_g} \right) \quad (16)$$

(Note that equation (16) is equivalent to $C_1(\Sigma)$ equation (9) provided $s = p$ i.e, Σ is of full rank) where λ_a and λ_g are the arithmetic mean and geometric mean of eigenvalues of Σ . Hence we interpret complexity as the log ratio between the arithmetic and the geometric mean of the eigenvalues of Σ . It measures how unequal the eigenvalues of Σ are, and incorporates the two simplest scalar measures of multivariate scatter, namely the trace and the determinant into a single function.

In general high values of complexity indicate a high interaction between variables and a low value of complexity indicates less interaction between variables. The minimum of $C_1(\Sigma)$ corresponds to the least complex structure. In other words $C_1(\Sigma)$ approaches zero as Σ approaches the identity matrix.

We can therefore say that orthogonal designs or linear models with no collinearity are the least complex therefore the identity matrix is the only matrix for which the complexity vanishes, otherwise $C_1(\Sigma) > 0$ (Bozdogan [11]).

We note that the correlation matrix can also be used to describe complexity. If we wish to show the interdependences (correlations) among variables we can transform the covariances to a correlation matrix.

1.2.3 PDI and Complexity Measure

PDI as defined earlier gives us the number of independent sources of variation in a portfolio. While the covariance complexity formula $C_1(\Sigma)$ carried out on the covari-

ance matrix of portfolios measures the degree of dependences among the securities that make up the portfolio.

The proposition here is that PDI and the complexity measure are two sides of the same coin. While PDI counts the number of independent factors in the portfolio complexity measure assigns a number to the degree of dependences among the dependent factors in a portfolio.

Intuitively we expect that a portfolio with a high PDI will consequently have a low Complexity measure. That is, as the number of independent factors driving variation in a portfolio returns increase the degree of dependency of all factors in the portfolio will decrease and vice versa.

A closer look may reveal that this may not necessarily be the case, as an alternative argument will be that number of independent sources of variation in a portfolio and degree of dependency among dependent sources of variation in a portfolio may not give the same rankings when applied to a set of portfolios. For example, if we have two portfolios A and B that have exactly two independent factors accounting for the variation in the portfolio and contains a total of 10 factors. PDI will return a value close to two for both portfolio but these may not necessarily mean that the complexity measure for the two portfolios will be the same. If the degree of dependence among the dependent factors in the two portfolios is different then we expect that the complexity measure for the two portfolios will be different.

Since the complexity measure like PDI is independent of the market index and going by the argument that increases in complexity of a portfolio should translate to decreases in its PDI, we rank a set of portfolios according to their PDI and complexity measure and discussed the result in the next section.

2 DIVERSIFICATION AND COMPLEXITY MEASURE OF PORTFOLIOS USING JSEDATA

Using JSEDATA used in the last analysis we formed six portfolios of shares in the following ways. The securities used are from the top 50 shares. The market weights

used are security weights at the end of the period (JULY 2009). These weights are kept constant for the whole period.

- Portfolio of randomly selected shares across sectors with market weights (Portfolio 1)
- Portfolio of randomly selected shares across sectors with equal weights (Portfolio 2)
- Portfolio of financial securities market weights (Portfolio 3)
- Portfolio of financial securities equal weights (Portfolio 4)
- Portfolio of Resources securities market weights (Portfolio 5)
- Portfolio of Resources securities equal weights (Portfolio 6)

The rationale is to select portfolios which we expect to have different diversification index going by the PDI measure and compare the ranking of the portfolio using the PDI measure to the ranking using the Complexity measure.

The limitation is that we had to construct dummy portfolios as we do not have access to data on weight of actual (real life) portfolios.

2.1 Methodology

The portfolios were formed by multiplying the rebalanced weights³ with the return of the shares over the period of interest. Then we conducted principal component analysis on the variance covariance matrix of the respective portfolios using an R statistical package. We calculated the PDI of each portfolio using the equation (15) for PDI given above. Equation (16) was used to calculate the complexity measure for the variance covariance matrix of each portfolio. The table 5 shows the results. First we compare complexity measure ranking to diversification ranking (as measured by PDI) of the portfolio. We then draw conclusions on the relationship between the

³Rebalanced weights were calculated by adding up the percentage market capitalization of each stock and rebalancing it such that the total portfolio weight sum up to one.

complexity of the portfolio and its measure of concentration. In general it is expected that diversification measured by PDI and Concentration measured by the formula given below should move in opposite directions.

$$Concentration = \sum_{i=1}^N w_i^2 \quad (17)$$

where w_i is the weight of share i in the index.

Table 5:DUMMY PORTFOLIO COMPOSITION AND CONCENTRATION

ACROSS SECTORS				
	Market Weight(1)		Equal Weight(2)	
Shares	Weight	$Weight^2$	Weight	$Weight^2$
MSM	0.03	0.0009	0.11	0.0121
APN	0.06	0.0036	0.11	0.0121
CFR	0.24	0.0576	0.11	0.0121
SHF	0.04	0.0016	0.11	0.0121
TBS	0.06	0.0036	0.11	0.0121
NTC	0.04	0.0016	0.11	0.0121
SAB	0.07	0.0049	0.11	0.0121
BVT	0.08	0.0064	0.11	0.0121
SOL	0.36	0.1296	0.11	0.0121
	Conc	0.2098	Conc	0.1089
FINANCIAL SECURITIES				
	Market Weight(3)		Equal Weight(4)	
Shares	Weight	$Weight^2$	Weight	$Weight^2$
ABL	0.07	0.0049	0.143	0.0205
ASA	0.25	0.0625	0.143	0.0205
FSR	0.28	0.0784	0.143	0.0205
INP	0.07	0.0049	0.143	0.0205
NED	0.19	0.0361	0.143	0.0205
RMH	0.1	0.01	0.143	0.0205
SBK	0.04	0.0016	0.143	0.0205
	Conc	0.1984	Conc	0.1431
RESOURCES SECURITIES				
	Market Weight(5)		Equal Weight(6)	
Shares	Weight	$Weight^2$	Weight	$Weight^2$
AGL	0.28	0.0784	0.11	0.0121
AMS	0.12	0.0144	0.11	0.0121
ARI	0.02	0.0004	0.11	0.0121
ANG	0.07	0.0049	0.11	0.0121
BIL	0.33	0.1089	0.11	0.0121
GFI	0.05	0.0025	0.11	0.0121
HAR	0.02	0.0004	0.11	0.0121
IMP	0.08	0.0064	0.11	0.0121
LON	0.03	0.0009	0.11	0.0121
	Conc	0.2172	Conc	0.1089

Conc stands for concentration

Table 5:PDI, COMPLEXITY AND CONCENTRATION RANKINGS FOR
PORTFOLIOS

	PDI RANKING (INCREASING)	COMPLEXITY RANKING (DECREASING)	CONCENTRATION RANKING (DECREASING)
1	Portfolio 3 (1.862212)	Portfolio 5 (18.29774)	Portfolio 5 (0.2172)
2	portfolio 1 (1.867859)	Portfolio 1 (16.86140)	Portfolio 1 (0.2098)
3	portfolio 5 (2.304017)	Portfolio 3 (11.44623)	Portfolio 3 (0.1984)
4	Portfolio 4 (2.485562)	Portfolio 4 (6.704973)	Portfolio 4 (0.1431)
5	Portfolio 6 (3.534385)	Portfolio 6 (3.534385)	Portfolio 6 (0.1089)
6	Portfolio 2 (4.1714886)	Portfolio 2 (3.496848)	Portfolio 2 (0.1089)

Table 6:RANK CORRELATION MATRIX OF PORTFOLIOS' PDI,
COMPLEXITY and CONC

	PDI	COM	CONC
PDI	1		
COM	-0.7714	1	
CONC	-0.7714	1	1

2.2 Findings

- As far as pairwise comparison is concerned between market weighted portfolios and the corresponding equally weighted portfolio (1 and 2, 3 and 4, 5 and 6) all the rankings agree. The equally weighted portfolio is less concentrated, more diversified and less complex than their market weighted counterparts.
- Although the Concentration and PDI tend to move in opposite direction Complexity measure agrees perfectly in terms of ranking with concentration across all portfolios. The rank correlation value of one in table 6 highlights this finding.
- The PDI measure and the complexity measure agree for all the other portfolios except for portfolio 3 and 5 (resources and financial securities market weight respectively). This implies that the complexity measure and PDI measure for portfolios do not always agree as argued earlier.
- For equally weighted portfolios the PDI and the complexity rankings agree. This suggests that provided the weights of securities in a portfolio is uniformly distributed an increase in diversification results in lower complexity. On the other

hand PDI and complexity ranking do not agree when market weights are used. Hence the less than perfect negative rank correlation between complexity and PDI on one hand and complexity and concentration on the other hand is as a result of market weighted portfolios.

2.3 Implications of findings

We could conclude from the first point above, that market weighted portfolios on the JSE tend to be less diversified and more complex (i.e has higher level of dependency among shares) than their equally weighted counterparts. This is due to the concentration in the JSE where resources shares dominate the market (Concentration is generally prevalent in emerging markets [12]).

We also note that if security weights are uniformly distributed in a portfolio, our initial proposition that diversification is negatively correlated with complexity measure is true for the set of data used in this analysis. This is however not the case when portfolio weights moves away from uniform distribution. Looking at the concentration ranking we note that portfolio 5 has the highest concentration of all the portfolios. Though portfolio 5 is more diversified than portfolio 1 and 3 it is more complex than the latter because of the effect of the distribution of its weights. It should however be noted that in this analysis portfolio weights are not real.

2.4 Conclusion

In stock market analysis, ICOMP select models that explains as much variation in securities returns as AIC type criteria. However, in contrast to AIC type criteria it selects models with more stable parameters by penalizing the interdependencies among the explanatory variables. Using the penalty term of ICOMP to measure complexity in the covariance matrix of a portfolio gives us a new way of looking at the diversification of a portfolio. While PDI counts the number of independent sources of variation in a portfolio, we could use complexity measure to quantify the level of interdependency among the securities in the portfolio. We found (within the

scope of data collected for this mini thesis) that provided the securities weights are evenly distributed a portfolio with a high number of independent sources of variation has low interdependency within its shares.

References

- [1] AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. 267–281. In *Petrov, BN and F. Csaki. Second International Symposium of Information Theory. Akademiai Kiado. Budapest, Hungary.*
- [2] BELSLEY, D., KUH, E., AND WELSCH, R. (1980). *Regression diagnostics.* Wiley New York.
- [3] BODIE, Z., KANE, A., AND MARCUS, A. (2006). *Investments.* .
- [4] BOX, G. (1976). Science and statistics. *Journal of the American Statistical Association* 71, 356 791–799.
- [5] BOYCE, D., FARHI, R., AND WEISCHEDEL, R. (1974). *Optimal subset selection: Multiple regression, interdependence and optimal network algorithms.* Springer-Verlag New York, Inc. Secaucus, NJ, USA.
- [6] BOZDOGAN, H. (1988). ICOMP: A new model selection criterion. *Classification and related methods of data analysis* 599–608.
- [7] BOZDOGAN, H. (1988). Theory and applications of information-theoretic measure of complexity (ICOMP) as a new model selection criterion. *Unpublished research report, the Institute of Statistical mathematics, Tokyo, Japan, and the Department of mathematics, University of Virginia, Charlottesville, VA,.*
- [8] BOZDOGAN, H.(1990). On the information-based measure of covariance complexity and its application to the evaluation of multivariate linear models. *Communications in statistics. Theory and methods* 19, 1 221–278.
- [9] BOZDOGAN, H. (1994). Mixture-model cluster analysis using a new informational complexity and model selection criteria. *Multivariate Statistical Modeling* 2 69–113.
- [10] BOZDOGAN, H. (2000). Akaike’s information criterion and recent developments in information complexity. *Journal of Mathematical Psychology* 44, 1.

- [11] BOZDOGAN, H. (2003). Intelligent statistical data mining with information complexity and genetic algorithms. *Proceeding of JISS 2003, Lisbonne 2* 15–56.
- [12] BRADFIELD, D., DUGMORE, B., AND GOPI, Y. (2006). A new measure of portfolio diversification. *Cadiz Report* 1–3.
- [13] CRAMER, H. (1946). *Mathematical Methods of statistics, 1946*. Department of Mathematical SU.
- [14] KULLBACK, S., AND LEIBLER, R. (1951). On information and sufficiency. *The Annals of Mathematical Statistics* 22, 1.
- [15] LINHART, H., AND ZUCCHINI, W. (1986) *Model selection*. Wiley New York et al..
- [16] MANTEL, N. (1970). Why stepdown procedures in variable selection. *Technometrics* 12, 3.
- [17] MARKOWITZ, H. (1952). Portfolio selection. *The Journal of Finance* 7, 1 77–91.
- [18] RAO, C. (1945). Information and accuracy attainable in the estimation of statistical . *Bull. Calcutta math. Soc* 37,81.
- [19] RAO, C., AND ANSCOMBE, F. (1949). Sufficient statistics and minimum variance estimates. In *Proceedings of the Cambridge Philosophical Society* vol. 45, pp. 45,213.
- [20] RAO, C., AND WISHART, J. (1947). Minimum variance and the estimation of several parameters. In *Proceedings of the Cambridge Philosophical Society* vol. 43, pp. 43,280.
- [21] RUDIN, A., AND MORGAN, S. (2006). A Portfolio Diversification Index. *The Journal of Portfolio Management* 32, 2 81–89.
- [22] SCHWARZ, G. (1978). Estimating the dimension of a model. *The annals of statistics* 6, 2.
- [23] SILVEY, S. (1969) Multicollinearity and imprecise estimation. *Journal of the Royal Statistical Society. Series B (Methodological)* 539–552.

- [24] SOKAL, R., AND ROHLF, F. (1981). *Biometry* (2nd edn). *New York: WH Feeman and Company* 668.
- [25] SPEED, F., HOCKING, R., AND HACKNEY, O. (1978). Methods of analysis of linear models with unbalanced data. *Journal of the American Statistical Association* 73, 361.
- [26] THIART, C. (1990). Collinearity and consequences for estimation : a stimulation study. *Unpublished Thesis (M.Sc. (Mathematical Statistics))–University of Cape Town*.
- [27] TROSKIE, C. G., AND CLARK, A. E. (2006). Regression and ICOMP - A Simulation Study. *Communications in Statistics* 35, 3.
- [28] VAN EMDEN, M. (1971). *An analysis of complexity*. Mathematisch Centrum.
- [29] WILKINSON, L. (1989). *The system for statistics*. Evanston, IL: Systat.

3 Appendix

Table of model selected by AIC, BIC and ICOMP using the all possible regression method. The table also displays the R^2 value for each model. The models include intercept term.

ALL POSSIBLE REGRESSION SELECTED MODELS FOR AIC, BIC AND ICOMP						
	AIC		BIC		ICOMP	
	R^2	MODEL	R^2	MODEL	R^2	MODEL
ABL	0.45	DJT + J200 + J203 + J835	0.43	J835	0.46	DJT + GOLD + J203 + J835 + JSEEX + R153
ACL	0.17	GOLD + J835 + R153	0.17	GOLD + J835 + R153	0.16	GOLD + J835 + JSEEX + R153
AEG	0.40	DJT+J200 + J203 + JSEEX + R153	0.38	J200 + J203 + R153	0.27	DJT + J211 + JSEEX + R153
AGL	0.85	GOLD + J210 + SP500	0.85	GOLD + J210 + SP500	0.85	GOLD + J210 + SP500
AMS	0.71	DJT + GOLD + J210 + SP500	0.71	GOLD + J210 + SP500	0.71	DJT + GOLD + J210 + J835 + JSEEX + SP500
ASA	0.71	GOLD + J200 + J203 + J210	0.67	J835	0.69	GOLD + J211 + J835 + JSEEX + R153
ARI	0.46	DJT + J210 + JSEEX + SP500	0.44	J210 + JSEEX	0.45	DJT + J210 + JSEEX
ANG	0.51	GOLD+J203+J210+J211+J835+SP500	0.48	GOLD + J210 + SP500	0.48	GOLD + J210 + SP500
APN	0.19	J200 + J203 + J835	0.16	J835	0.18	GOLD + J835 + JSEEX + R153 + SP500
BIL	0.72	DJT + J200 + J203 + J835	0.79	GOLD + J200 + J211	0.80	DJT + GOLD + J210 + J211 + J835
CFR	0.67	J200 + SP500	0.67	J200 + SP500	0.67	GOLD + J203 + R153 + SP500
DDT	0.54	GOLD + J211 + JSEEX + R 153 + SP500	0.51	J211 + JSEEX	0.54	GOLD + J211 + J835 + JSEEX + R153 + SP500
DSY	0.31	J203 + J210 + J835 + JSEEX + SP500	0.30	J203 + J210 + J835 + SP500	0.26	J210 + J211 + J835 + JSEEX + R153 + SP500
FSR	0.90	J200 + J210 + J211 + J835	0.89	J835	0.89	J835
GFI	0.55	GOLD + J210 + JSEEX + SP500	0.54	GOLD + J210 + SP500	0.55	GOLD + J210 + JSEEX + SP500
GRT	0.43	J200 + J203 + J835 + R 153	0.42	J200 + J203 + R153	0.41	GOLD + J203 + J835 + JSEEX + R153
HAR	0.52	GOLD + J210 + JSEEX + SP500	0.51	GOLD + J210 + SP500	0.52	GOLD + J210 + JSEEX + R153 + SP500
IMP	0.67	GOLD + J200 + J211 + SP500	0.67	GOLD + J200 + J211	0.67	GOLD + J200 + J211 + JSEEX + R153 + SP500
IPL	0.45	GOLD + J200 + J203 + JSEEX + R 153	0.40	J200 + J203	0.42	GOLD + J203 + J210 + JSEEX + R153 + SP500
INP	0.51	DJT + J200 + J210 + R 153 + SP500	0.49	J200 + J210	0.51	DJT + GOLD + J200 + J210 + R153 + SP500
LBH	0.30	J203 + J210 + J211 + J835 + SP500	0.30	J203 + J210 + J211 + J835 + SP500	0.26	J203 + J210 + J835 + SP500
LBT	0.35	J835 + JSEEX + R 153 +SP500	0.35	J835 + JSEEX + R153 + SP500	0.35	GOLD + J210 + J211 + J835 + JSEEX + R153 + SP500
LON	0.63	J200 + J203 + J211 + J835 + JSEEX + R 153	0.62	J200 + J203 + J211 + J835 + R153 + SP500	0.58	J203 + J211 + J835 + JSEEX + R153 + SP500
MSM	0.35	J200 + J203 + J211	0.35	J200 + J203 + J211	0.28	DJT + J200 + J211 + J835 + R153
MTN	0.35	DJT + J200 + J203 + J211 + JSEEX + R 153	0.51	J211 + JSEEX	0.54	DJT + J203 + J211 + JSEEX + R153 + SP500
MUR	0.33	J200 + J203 + J210 + J835 + R 153 + SP500	0.28	J200 + J203	0.20	J203 + JSEEX
NPN	0.52	DJT + GOLD + J211 + JSEEX + R 153 + SP500	0.48	J211 + R153	0.52	DJT + GOLD + J211 + J835 + JSEEX + R153 + SP500
NED	0.62	J203 + J210 + J835	0.62	J203 + J210 + J835	0.61	J210 + J211 + J835 + JSEEX + R153
NTC	0.36	J200 + J203 + JSEEX	0.35	J200 + J203	0.35	J200 + J203
OML	0.66	J203 + J210 + J211 + SP500	0.66	J203 + J210 + J211 + SP500	0.63	J203 + J835 + SP500
PIK	0.28	J203 + J210 + JSEEX + SP500	0.25	J203 + J210	0.27	GOLD + J211 + J835 + JSEEX + R153 + SP500
PPC	0.32	GOLD + J200 + J203 + JSEEX + R 153	0.27	J200 + J203	0.30	GOLD + J203 + J835 + JSEEX + R153
REM	0.41	GOLD + J203 + J210 + J211 + SP500	0.41	GOLD + J203 + J210 + J211 + SP500	0.39	GOLD + J210 + J211 + J835 + SP500
RMH	0.90	J835	0.90	J835	0.90	DJT + GOLD + J835 + JSEEX + R153
SAB	0.53	J211 + JSEEX + R 153	0.52	J211 + JSEEX	0.52	GOLD + J211 + JSEEX + R153
SLM	0.57	DJT + J200 + J210 + J211 + R153	0.56	J203 + J210 + J211	0.54	DJT + J203 + J210 + R153
SAP	0.40	DJT+J200 + J203 + J835 + R 153	0.37	J203 + J835 + SP500	0.37	J203 + J835 + R153 + SP500
SOL	0.63	J210 + R153	0.62	J210	0.63	J210 + JSEEX + R153
SHP	0.40	J200 + J203 + J211 + JSEEX + SP500	0.38	J200 + J211 + JSEEX + SP500	0.37	J210 + J211 + JSEEX + SP500
SBK	0.92	GOLD + J200 + J203 + J835 + SP500	0.91	J835	0.92	GOLD + J211 + J835 + JSEEX + R153 + SP500
SHF	0.51	J200 + J203 + J211 + JSEEX + R 153	0.49	J200 + J203 + J211 + R153	0.43	J210 + J211 + JSEEX + R153
BVT	0.63	GOLD + J200 + J203 + J210	0.63	GOLD + J200 + J203 + J210	0.60	DJT + GOLD + J203 + J210 + JSEEX + R153
TBS	0.42	J211 + JSEEX + SP500	0.42	J211 + JSEEX + SP500	0.42	GOLD + J211 + J835 + JSEEX + SP500
TRU	0.36	J200 + J203 + J210 + R 153	0.34	J200 + J203 + R153	0.32	GOLD + J211 + J835 + JSEEX + R153

Table of model selected by AIC for forward backward and stepwise procedures. The table also displays the R^2 value for each model. The models include intercept term.

MODEL SELECTED UNDER AIC FOR FORWARD, BACKWARD AND STEPWISE PROCEDURES						
STOCK	Forward		Backward		Stepwise	
	R^2	MODEL	R^2	MODEL	R	MODEL
ABL	0.4312	J835	0.4552	J203+J835+GOLD+JSEEX+R153+DJT	0.4469	J835
ACL	0.1623	J203+GOLD+J835+R153	0.1676	J200+J211+J210+GOLD+R153	0.1692	GOLD+J835+R153
AEG	0.27	J211+JSEEX	0.4017	J203+J200+JSEEX+R153+DJT	0.4017	J211+JSEEX
AGL	0.8511	J210+GOLD+SP500	0.8511	J210+GOLD+SP500	0.8511	J210+GOLD+SP500
AMS	0.7108	J210+GOLD+SP500+DJT	0.7108	J210+GOLD+SP500+DJT	0.7108	J210+GOLD+SP500+DJT
ASA	0.6892	J835+R153+GOLD	0.7094	J203+J200+J835+J210+GOLD	0.7094	J835+R153+GOLD
ARI	0.4553	J210+JSEEX+DJT+SP500	0.4553	J210+JSEEX+DJT+SP500	0.4553	J210+JSEEX+DJT+SP500
ANG	0.4813	J210+GOLD+SP500	0.5129	J203+J835+J211+J210+GOLD+SP500	0.5129	J210+GOLD+SP500
APN	0.1808	J835+SP500	0.1894	J203+J200+J835	0.1894	J835+SP500
BIL	0.7868	J210+GOLD	0.7989	J200+J211+J210+GOLD+DJT	0.7195	J210+GOLD
CFR	0.6664	SP500+J200	0.6664	SP500+J200	0.6664	SP500+J200
DDT	0.538	J211+JSEEX+GOLD+SP500+R153	0.538	J211+JSEEX+GOLD+SP500+R153	0.538	J211+JSEEX+GOLD+SP500+R153
DSY	0.2575	J211+DJT+J210	0.308	J203+J835+J210+JSEEX+SP500	0.308	J211+DJT+J210
FSR	0.8998	J835	0.8954	J200+J835+J211+J210	0.8954	J835
GFI	0.5468	J210+GOLD+SP500+JSEEX	0.5473	J203+J211+J210+GOLD+JSEEX+SP500	0.5468	J210+GOLD+SP500+JSEEX
GRT	0.4107	R153+J835+J211	0.4322	J203+J200+J835+R153	0.4322	R153+J835+J211
HAR	0.5246	J210+SP500+GOLD+JSEEX	0.5246	J210+SP500+GOLD+JSEEX	0.5246	J210+SP500+GOLD+JSEEX
IMP	0.6619	J210+GOLD+SP500	0.6696	J200+J211+GOLD+SP500	0.6696	J210+GOLD+SP500
IPL	0.3845	J835+J211+R153+GOLD	0.4461	J203+J200+GOLD+JSEEX+R153	0.4461	J835+J211+R153+GOLD
INP	0.4833	J211+J835+DJT+R153	0.5107	J200+J210+R153+DJT+SP500	0.5107	J211+J835+DJT+R153
LBH	0.3017	J211+J210+J203+J835+SP500	0.3017	J203+J835+J211+J210+SP500	0.3017	J211+J210+J203+J835+SP500
LBT	0.3568	J200+JSEEX+SP500+R153+J835+J211	0.353	J835+JSEEX+R153+SP500	0.353	JSEEX+SP500+J835+R153
LON	0.5578	J210+SP500+J835	0.6273	J203+J200+J835+J211+JSEEX+R153+SP500	0.6273	J210+SP500+J835
MSM	0.3553	J835+J211+J200+J203+J210	0.3553	J203+J200+J835+J211+J210	0.3477	J211+J200+J203
MTN	0.5646	J211+JSEEX+J203+J200+R153+SP500+DJT	0.5646	J203+J200+J211+JSEEX+R153+DJT+SP500	0.5646	J211+JSEEX+J203+J200+R153+SP500+DJT
MUR	0.3308	J203+J200+J210+SP500+J835+R153	0.3308	J203+J200+J835+J210+R153+SP500	0.3308	J203+J200+J210+SP500+J835+R153
NPN	0.5027	J211+R153+JSEEX+GOLD	0.5204	J211+GOLD+JSEEX+R153+DJT+SP500	0.5204	J211+R153+JSEEX+GOLD
NED	0.6129	J835+J211+J210	0.6129	J835+J211+J210	0.6183	J835+J211+J210
NTC	0.3512	J211+R153+J203+J200+JSEEX	0.358	J203+J200+JSEEX	0.358	J203+J200+JSEEX
OML	0.6628	SP500+J835+J203+J210+J211	0.6624	J203+J211+J210+SP500	0.6624	SP500+J203+J210+J211
PIK	0.2745	J835+J211+J200+J203	0.2818	J203+J200+J211+JSEEX+SP500	0.2764	J835+J211+J200+J203
PPC	0.2802	J835+R153+J211	0.3247	J203+J200+GOLD+JSEEX+R153	0.3247	J835+R153+J211
REM	0.4079	SP500+GOLD+J835+J211+J203+J210	0.412	J203+J211+J210+GOLD+SP500	0.412	SP500+GOLD+J211+J203+J210
RMH	0.8971	J835	0.8971	J835	0.8971	J835
SAB	0.5262	J211+JSEEX+R153	0.5262	J211+JSEEX+R153	0.5262	J211+JSEEX+R153
SLM	0.501	J835+J211+R153	0.501	J835+J211+R153	0.5718	J835+J211+R153
SAP	0.4009	J203+DJT+J835+J200+R153	0.4009	J203+J200+J835+R153+DJT	0.4009	J203+DJT+J835+J200+R153
SOL	0.6315	J210+R153	0.6315	J210+R153	0.6315	J210+R153
SHP	0.3967	J211+J200+J203+SP500+JSEEX	0.3967	J211+J200+J203+SP500+JSEEX	0.3967	J211+J200+J203+SP500+JSEEX
SEK	0.9164	J835+R153+GOLD	0.9194	J203+J200+J835+GOLD+SP500	0.9194	J835+R153+GOLD
SHF	0.4259	J211	0.5053	J203+J200+J211+JSEEX+R153	0.5053	J211
BVT	0.5496	J211+J835+GOLD+R153	0.6293	J203+J200+J210+GOLD	0.6293	J211+J835+GOLD+R153
TBS	0.4243	J211+JSEEX+SP500	0.4243	J211+JSEEX+SP500	0.4243	J211+JSEEX+SP500
TRU	0.3585	J835+R153+J211+J200+J203	0.3595	J203+J200+J835+R153+DJT	0.3593	R153+J211+J200+J203

Table of model selected by BIC for forward backward and stepwise procedures. The table also displays the R^2 value for each model. The models include intercept term.

MODEL SELECTED UNDER BIC FOR FORWARD, BACKWARD AND STEPWISE PROCEDURES						
STOCK	Forward		Backward		Stepwise	
	R^2	MODEL	R^2	MODEL	R^2	MODEL
ABL	0.43	J835	0.456	J203+J835+GOLD+JSEEX+R153+DJT	0.43	J835
ACL	0.11	J203+GOLD	0.17	J200+J211+J210+GOLD+R153	0.16	GOLD+J835+R153
AEG	0.27	J211+JSEEX	0.401	J203+J200+JSEEX+R153+DJT	0.27	J211+JSEEX
AGL	0.85	J210+GOLD+SP500	0.85	J210+GOLD+SP500	0.85	J210+GOLD+SP500
AMS	0.71	J210+GOLD+SP500	0.71	J210+GOLD+SP500+DJT	0.71	J210+GOLD+SP500+DJT
ASA	0.68	J835	0.71	J203+J200+J835+J210+GOLD	0.68	J835+R153+GOLD
ARI	0.44	J210+JSEEX	0.46	J210+JSEEX+DJT+SP500	0.45	J210+JSEEX+DJT+SP500
ANG	0.48	J210+GOLD+SP500	0.51	J203+J835+J211+J210+GOLD+SP500	0.48	J210+GOLD+SP500
APN	0.16	J835	0.19	J203+J200+J835	0.18	J835+SP500
BIL	0.79	J210+GOLD	0.79	J200+J211+J210+GOLD+DJT	0.78	J210+GOLD
CFR	0.67	SP500+J200	0.67	SP500+J200	0.66	SP500+J200
DDT	0.51	J211+JSEEX	0.54	J211+JSEEX+GOLD+SP500+R153	0.53	J211+JSEEX+GOLD+SP500+R153
DSY	0.22	J211	0.31	J203+J835+J210+JSEEX+SP500	0.25	J211+DJT+J210
FSR	0.89	J835	0.89	J200+J835+J211+J210	0.88	J835
GFI	0.54	J210+GOLD+SP500	0.55	J203+J211+J210+GOLD+JSEEX+SP500	0.54	J210+GOLD+SP500+JSEEX
GRT	0.39	R153+J835	0.43	J203+J200+J835+R153	0.41	R153+J835+J211
HAR	0.51	J210+SP500+GOLD	0.52	J210+SP500+GOLD+JSEEX	0.52	J210+SP500+GOLD+JSEEX
IMP	0.66	J210+GOLD+SP500	0.67	J200+J211+GOLD+SP500	0.66	J210+GOLD+SP500
IPL	0.38	J835+J211+R153	0.45	J203+J200+GOLD+JSEEX+R153	0.38	J835+J211+R153+GOLD
INP	0.47	J211+J835	0.51	J200+J210+R153+DJT+SP500	0.48	J211+J835+DJT+R153
LBH	0.13	J211	0.30	J203+J835+J211+J210+SP500	0.30	J211+J210+J203+J835+SP500
LBT	0.31	J200+JSEEX+SP500	0.35	J835+JSEEX+R153+SP500	0.35	JSEEX+SP500+J835+R153
LON	0.56	J210+SP500+J835	0.64	J203+J200+J835+J211+JSEEX+R153+SP500	0.55	J210+SP500+J835
MSM	0.15	J835	0.36	J203+J200+J835+J211+J210	0.34	J211+J200+J203
MTN	0.51	J211+JSEEX	0.57	J203+J200+J211+JSEEX+R153+DJT+SP500	0.56	J211+JSEEX+J203+J200+R153+SP500+DJT
MUR	0.28	J203+J200	0.33	J203+J200+J835+J210+R153+SP500	0.33	J203+J200+J210+SP500+J835+R153
NPN	0.48	J211+R153	0.52	J211+GOLD+JSEEX+R153+DJT+SP500	0.50	J211+R153+JSEEX+GOLD
NED	0.57	J835	0.61	J835+J211+J210	0.61	J835+J211+J210
NTC	0.21	J211+R153	0.36	J203+J200+JSEEX	0.35	J203+J200+JSEEX
OML	0.63	SP500+J835+J203	0.66	J203+J211+J210+SP500	0.66	SP500+J203+J210+J211
PIK	0.24	J835+J211	0.28	J203+J200+J211+JSEEX+SP500	0.27	J835+J211+J200+J203
PPC	0.22	J835	0.32	J203+J200+GOLD+JSEEX+R153	0.28	J835+R153+J211
REM	0.37	SP500+GOLD+J835	0.41	J203+J211+J210+GOLD+SP500	0.41	SP500+GOLD+J211+J203+J210
RMH	0.89	J835	0.89	J835	0.89	J835
SAB	0.52	J211+JSEEX	0.53	J211+JSEEX+R153	0.52	J211+JSEEX+R153
SLM	0.49	J835+J211	0.50	J835+J211+R153	0.50	J835+J211+R153
SAP	0.34	J203+DJT	0.40	J203+J200+J835+R153+DJT	0.40	J203+DJT+J835+J200+R153
SOL	0.62	J210	0.63	J210+R153	0.63	J210+R153
SHP	0.34	J211+J200+J203	0.39	J211+J200+J203+SP500+JSEEX	0.39	J211+J200+J203+SP500+JSEEX
SEK	0.91	J835	0.92	J203+J200+J835+GOLD+SP500	0.91	J835+R153+GOLD
SHF	0.43	J211	0.51	J203+J200+J211+JSEEX+R153	0.42	J211
BVT	0.51	J211+J835	0.63	J203+J200+J210+GOLD	0.54	J211+J835+GOLD+R153
TBS	0.42	J211+JSEEX+SP500	0.42	J211+JSEEX+SP500	0.42	J211+JSEEX+SP500
TRU	0.31	J835+R153+J211	0.36	J203+J200+J835+R153+DJT	0.35	R153+J211+J200+J203

R Code

FORWARD, BACKWARD, STEPWISE AND ALL POSSIBLE REGRESSION PROCEDURE FOR AIC

```
setwd("Q:")
StockData<-read.csv(file="Top50Monthly(2).csv",header=TRUE);
n=nrow(StockData);
m=ncol(StockData);
J203=StockData[1:n,45];
J200=StockData[1:n,46];
J835=StockData[1:n,47];
J211=StockData[1:n,48];
J210=StockData[1:n,49];
GOLD=StockData[1:n,50];
JSEEX=StockData[1:n,51];
R153=StockData[1:n,52];
DJT=StockData[1:n,53];
SP500=StockData[1:n,54];
k=m-11
fitforward=rep(0,k);
fitbackward=rep(0,k);
fitStepwise=rep(0,k);

yvect=c('ABL','ACL','AEG','AGL','AMS','ASA','ARI','ANG','APN','BIL','CFR','DDT','DSY','FSR','GFI','GRT','HAR','IMP','IPL','INP','LBH','LBT','LON','MSM','MTN','MUR','NPN','NED','NTC','OML','PIK','PPO')

#Forward procedure
for (i in 1:k){
fitforward[i]=step(lm(StockData[1:n,i]~1),StockData[1:n,i]~J203+J200+J835+J211+J210+GOLD+JSEEX+R153+DJT+SP500, direction = "forward")
}

#ANOVA for forward Procedure
summary(lm(StockData[1:n,1] ~ J835));
summary(lm(StockData[1:n,2] ~ J203+GOLD+J835+R153));
summary(lm(StockData[1:n,3] ~ J211+JSEEX));
summary(lm(StockData[1:n,4] ~ J210+GOLD+SP500));
summary(lm(StockData[1:n,5] ~ J210+GOLD+SP500+DJT));
summary(lm(StockData[1:n,6] ~ J835+R153+GOLD));
summary(lm(StockData[1:n,7] ~ J210+JSEEX+DJT+SP500));
summary(lm(StockData[1:n,8] ~ J210+GOLD+SP500));
summary(lm(StockData[1:n,9] ~ J835+SP500));
summary(lm(StockData[1:n,10] ~ J210+GOLD));
summary(lm(StockData[1:n,11] ~ SP500+J200));
summary(lm(StockData[1:n,12] ~ J211+JSEEX+GOLD+SP500+R153));
summary(lm(StockData[1:n,13] ~ J211+DJT+J210));
summary(lm(StockData[1:n,14] ~ J835));
summary(lm(StockData[1:n,15] ~ J210+GOLD+SP500+JSEEX));
summary(lm(StockData[1:n,16] ~ R153+J835+J211));
summary(lm(StockData[1:n,17] ~ J210+SP500+GOLD+JSEEX));
summary(lm(StockData[1:n,18] ~ J210+GOLD+SP500));
summary(lm(StockData[1:n,19] ~ J835+J211+R153+GOLD));
summary(lm(StockData[1:n,20] ~ J211+J835+DJT+R153));
summary(lm(StockData[1:n,21] ~ J211+J210+J203+J835+SP500));
summary(lm(StockData[1:n,22] ~ J200+JSEEX+SP500+R153+J835+J211));
summary(lm(StockData[1:n,23] ~ J210+SP500+J835));
summary(lm(StockData[1:n,24] ~ J835+J211+J200+J203+J210));
summary(lm(StockData[1:n,25] ~ J211+JSEEX+J203+J200+R153+SP500+DJT));
summary(lm(StockData[1:n,26] ~ J203+J200+J210+SP500+J835+R153));
summary(lm(StockData[1:n,27] ~ J211+R153+JSEEX+GOLD));
summary(lm(StockData[1:n,28] ~ J835+J211+J210));
summary(lm(StockData[1:n,29] ~ J211+R153+J203+J200+JSEEX));
summary(lm(StockData[1:n,30] ~ SP500+J835+J203+J210+J211));
summary(lm(StockData[1:n,31] ~ J835+J211+J200+J203));
summary(lm(StockData[1:n,32] ~ J835+R153+J211));
summary(lm(StockData[1:n,33] ~ SP500+GOLD+J835+J211+J203+J210));
summary(lm(StockData[1:n,34] ~ J835));
summary(lm(StockData[1:n,35] ~ J211+JSEEX+R153));
```

```

summary(lm(StockData[1:n,36] ~ J835+J211+R153));
summary(lm(StockData[1:n,37] ~ J203+DJT+J835+J200+R153));
summary(lm(StockData[1:n,38] ~ J210+R153));
summary(lm(StockData[1:n,39] ~ J211+J200+J203+SP500+JSEEX));
summary(lm(StockData[1:n,40] ~ J835+R153+GOLD));
summary(lm(StockData[1:n,41] ~ J211));
summary(lm(StockData[1:n,42] ~ J211+J835+GOLD+R153));
summary(lm(StockData[1:n,43] ~ J211+JSEEX+SP500));
summary(lm(StockData[1:n,44] ~ J835+R153+J211+J200+J203));

#Backward procedure
for (i in 1:k){
fitbackward[i]=step(lm(StockData[1:n,i] ~ J203+J200+J835+J211+J210+GOLD+JSEEX+R153+DJT+SP500),direction = "backward");
}

fitbackward;

#ANOVA for Backward Procedure
summary(lm(StockData[1:n,1] ~ J203+J835+GOLD+JSEEX+R153+DJT));
summary(lm(StockData[1:n,2] ~ J200+J211+J210+GOLD+R153));
summary(lm(StockData[1:n,3] ~ J203+J200+JSEEX+R153+DJT));
summary(lm(StockData[1:n,4] ~ J210+GOLD+SP500));
summary(lm(StockData[1:n,5] ~ J210+GOLD+SP500+DJT));
summary(lm(StockData[1:n,6] ~ J203+J200+J835+J210+GOLD));
summary(lm(StockData[1:n,7] ~ J210+JSEEX+DJT+SP500));
summary(lm(StockData[1:n,8] ~ J203+J835+J211+J210+GOLD+SP500));
summary(lm(StockData[1:n,9] ~ J203+J200+J835));
summary(lm(StockData[1:n,10] ~ J200+J211+J210+GOLD+DJT));
summary(lm(StockData[1:n,11] ~ SP500+J200));
summary(lm(StockData[1:n,12] ~ J211+JSEEX+GOLD+SP500+R153));
summary(lm(StockData[1:n,13] ~ J203+J835+J210+JSEEX+SP500));
summary(lm(StockData[1:n,14] ~ J200+J835+J211+J210));
summary(lm(StockData[1:n,15] ~ J203+J211+J210+GOLD+JSEEX+SP500));
summary(lm(StockData[1:n,16] ~ J203+J200+J835+R153));
summary(lm(StockData[1:n,17] ~ J210+SP500+GOLD+JSEEX));
summary(lm(StockData[1:n,18] ~ J200+J211+GOLD+SP500));
summary(lm(StockData[1:n,19] ~ J203+J200+GOLD+JSEEX+R153));
summary(lm(StockData[1:n,20] ~ J200+J210+R153+DJT+SP500));
summary(lm(StockData[1:n,21] ~ J203+J835+J211+J210+SP500));
summary(lm(StockData[1:n,22] ~ J835+JSEEX+R153+SP500));
summary(lm(StockData[1:n,23] ~ J203+J200+J835+J211+JSEEX+R153+SP500));
summary(lm(StockData[1:n,24] ~ J203+J200+J835+J211+J210));
summary(lm(StockData[1:n,25] ~ J203+J200+J211+JSEEX+R153+DJT+SP500));
summary(lm(StockData[1:n,26] ~ J203+J200+J835+J210+R153+SP500));
summary(lm(StockData[1:n,27] ~ J211+GOLD+JSEEX+R153+DJT+SP500));
summary(lm(StockData[1:n,28] ~ J835+J211+J210));
summary(lm(StockData[1:n,29] ~ J203+J200+JSEEX));
summary(lm(StockData[1:n,30] ~ J203+J211+J210+SP500));
summary(lm(StockData[1:n,31] ~ J203+J200+J211+JSEEX+SP500));
summary(lm(StockData[1:n,32] ~ J203+J200+GOLD+JSEEX+R153));
summary(lm(StockData[1:n,33] ~ J203+J211+J210+GOLD+SP500));
summary(lm(StockData[1:n,34] ~ J835));
summary(lm(StockData[1:n,35] ~ J211+JSEEX+R153));
summary(lm(StockData[1:n,36] ~ J835+J211+R153));
summary(lm(StockData[1:n,37] ~ J203+J200+J835+R153+DJT));
summary(lm(StockData[1:n,38] ~ J210+R153));
summary(lm(StockData[1:n,39] ~ J211+J200+J203+SP500+JSEEX));
summary(lm(StockData[1:n,40] ~ J203+J200+J835+GOLD+SP500));
summary(lm(StockData[1:n,41] ~ J203+J200+J211+JSEEX+R153));
summary(lm(StockData[1:n,42] ~ J203+J200+J210+GOLD));
summary(lm(StockData[1:n,43] ~ J211+JSEEX+SP500));
summary(lm(StockData[1:n,44] ~ J203+J200+J835+R153+DJT));

for (i in 1:k){

```

```

fitStepwise[i]=step(lm(StockData[1:n,i]~1),StockData[1:n,i]~J203+J200+J835+J211+J210+GOLD+JSEEX+R153+DJT+SP500, direction = "both")
}

fitStepwise;

#ANOVA for Stepwise Procedure
summary(lm(StockData[1:n,1] ~ J835));
summary(lm(StockData[1:n,2] ~ GOLD+J835+R153));
summary(lm(StockData[1:n,3] ~ J211+JSEEX));
summary(lm(StockData[1:n,4] ~ J210+GOLD+SP500));
summary(lm(StockData[1:n,5] ~ J210+GOLD+SP500+DJT));
summary(lm(StockData[1:n,6] ~ J835+R153+GOLD));
summary(lm(StockData[1:n,7] ~ J210+JSEEX+DJT+SP500));
summary(lm(StockData[1:n,8] ~ J210+GOLD+SP500));
summary(lm(StockData[1:n,9] ~ J835+SP500));
summary(lm(StockData[1:n,10] ~ J210+GOLD));
summary(lm(StockData[1:n,11] ~ SP500+J200));
summary(lm(StockData[1:n,12] ~ J211+JSEEX+GOLD+SP500+R153));
summary(lm(StockData[1:n,13] ~ J211+DJT+J210));
summary(lm(StockData[1:n,14] ~ J835));
summary(lm(StockData[1:n,15] ~ J210+GOLD+SP500+JSEEX));
summary(lm(StockData[1:n,16] ~ R153+J835+J211));
summary(lm(StockData[1:n,17] ~ J210+SP500+GOLD+JSEEX));
summary(lm(StockData[1:n,18] ~ J210+GOLD+SP500));
summary(lm(StockData[1:n,19] ~ J835+J211+R153+GOLD));
summary(lm(StockData[1:n,20] ~ J211+J835+DJT+R153));
summary(lm(StockData[1:n,21] ~ J211+J210+J203+J835+SP500));
summary(lm(StockData[1:n,22] ~ JSEEX+SP500+J835+R153));
summary(lm(StockData[1:n,23] ~ J210+SP500+J835));
summary(lm(StockData[1:n,24] ~ J211+J200+J203 ));
summary(lm(StockData[1:n,25] ~ J211+JSEEX+J203+J200+R153+SP500+DJT));
summary(lm(StockData[1:n,26] ~ J203+J200+J210+SP500+J835+R153));
summary(lm(StockData[1:n,27] ~ J211+R153+JSEEX+GOLD));
summary(lm(StockData[1:n,28] ~ J835+J211+J210));
summary(lm(StockData[1:n,29] ~ J203+J200+JSEEX));
summary(lm(StockData[1:n,30] ~ SP500+J203+J210+J211));
summary(lm(StockData[1:n,31] ~ J835+J211+J200+J203));
summary(lm(StockData[1:n,32] ~ J835+R153+J211));
summary(lm(StockData[1:n,33] ~ SP500+GOLD+J211+J203+J210));
summary(lm(StockData[1:n,34] ~ J835));
summary(lm(StockData[1:n,35] ~ J211+JSEEX+R153));
summary(lm(StockData[1:n,36] ~ J835+J211+R153));
summary(lm(StockData[1:n,37] ~ J203+DJT+J835+J200+R153));
summary(lm(StockData[1:n,38] ~ J210+R153));
summary(lm(StockData[1:n,39] ~ J211+J200+J203+SP500+JSEEX));
summary(lm(StockData[1:n,40] ~ J835+R153+GOLD));
summary(lm(StockData[1:n,41] ~ J211));
summary(lm(StockData[1:n,42] ~ J211+J835+GOLD+R153));
summary(lm(StockData[1:n,43] ~ J211+JSEEX+SP500));
summary(lm(StockData[1:n,44] ~ R153+J211+J200+J203));

#All possible
ABL=StockData[1:n,1];
ACL=StockData[1:n,2];
AEG=StockData[1:n,3];
AGL=StockData[1:n,4];
AMS=StockData[1:n,5];
ASA=StockData[1:n,6];
ARI=StockData[1:n,7];
ANG=StockData[1:n,8];
APN=StockData[1:n,9];
BIL=StockData[1:n,1];
CFR=StockData[1:n,11];
DDT=StockData[1:n,12];
DSY=StockData[1:n,13];
FSR=StockData[1:n,14];

```

```

GFI=StockData[1:n,15];
GRT=StockData[1:n,16];
HAR=StockData[1:n,17];
IMP=StockData[1:n,18];
IPL=StockData[1:n,19];
INP=StockData[1:n,20];
LBH=StockData[1:n,21];
LBT=StockData[1:n,22];
LON=StockData[1:n,23];
MSM=StockData[1:n,24];
MTN=StockData[1:n,25];
MUR=StockData[1:n,26];
NPN=StockData[1:n,27];
NED=StockData[1:n,28];
NTC=StockData[1:n,29];
OML=StockData[1:n,30];
PIK=StockData[1:n,31];
PPC=StockData[1:n,32];
REM=StockData[1:n,33];
RMH=StockData[1:n,34];
SAB=StockData[1:n,35];
SLM=StockData[1:n,36];
SAP=StockData[1:n,37];
SOL=StockData[1:n,38];
SHP=StockData[1:n,39];
SBK=StockData[1:n,40];
SHF=StockData[1:n,41];
BVT=StockData[1:n,42];
TBS=StockData[1:n,43];
TRU=StockData[1:n,44];
J203=StockData[1:n,45];
J200=StockData[1:n,46];
J835=StockData[1:n,47];
J211=StockData[1:n,48];
J210=StockData[1:n,49];
GOLD=StockData[1:n,50];
JSEEX=StockData[1:n,51];
R153=StockData[1:n,52];
DJT=StockData[1:n,53];
SP500=StockData[1:n,54];

#ABL
v=dredge(lm(ABL~J203+J200+J835+J211+J210+GOLD+JSEEX+R153+DJT+SP500),eval=TRUE,rank="AIC")
w=attributes(v)
k=Sys.getenv(w$formulas[1]);
k

#ACL
v=dredge(lm(ACL~J203+J200+J835+J211+J210+GOLD+JSEEX+R153+DJT+SP500),eval=TRUE,rank="AIC")
w=attributes(v)
k=Sys.getenv(w$formulas[1]);
k

#AEG
v=dredge(lm(AEG~J203+J200+J835+J211+J210+GOLD+JSEEX+R153+DJT+SP500),eval=TRUE,rank="AIC")
w=attributes(v)
k=Sys.getenv(w$formulas[1]);
k

#AGL
v=dredge(lm(AGL~J203+J200+J835+J211+J210+GOLD+JSEEX+R153+DJT+SP500),eval=TRUE,rank="AIC")
w=attributes(v)
k=Sys.getenv(w$formulas[1]);
k

#AMS
v=dredge(lm(AMS~J203+J200+J835+J211+J210+GOLD+JSEEX+R153+DJT+SP500),eval=TRUE,rank="AIC")
w=attributes(v)
k=Sys.getenv(w$formulas[1]);
k

```

```

#ASA
v=dredge(lm(ASA~J203+J200+J835+J211+J210+GOLD+JSEEX+R153+DJT+SP500),eval=TRUE,rank="AIC")
w=attributes(v)
k=Sys.getenv(w$formulas[1]);
k

#ARI
v=dredge(lm(ARI~J203+J200+J835+J211+J210+GOLD+JSEEX+R153+DJT+SP500),eval=TRUE,rank="AIC")
w=attributes(v)
k=Sys.getenv(w$formulas[1]);
k

#ANG
v=dredge(lm(ANG~J203+J200+J835+J211+J210+GOLD+JSEEX+R153+DJT+SP500),eval=TRUE,rank="AIC")
w=attributes(v)
k=Sys.getenv(w$formulas[1]);
k

#APN
v=dredge(lm(APN~J203+J200+J835+J211+J210+GOLD+JSEEX+R153+DJT+SP500),eval=TRUE,rank="AIC")
w=attributes(v)
k=Sys.getenv(w$formulas[1]);
k

#BIL
v=dredge(lm(BIL~J203+J200+J835+J211+J210+GOLD+JSEEX+R153+DJT+SP500),eval=TRUE,rank="AIC")
w=attributes(v)
k=Sys.getenv(w$formulas[1]);
k

#CFR
v=dredge(lm(CFR~J203+J200+J835+J211+J210+GOLD+JSEEX+R153+DJT+SP500),eval=TRUE,rank="AIC")
w=attributes(v)
k=Sys.getenv(w$formulas[1]);
k

#DDT
v=dredge(lm(DDT~J203+J200+J835+J211+J210+GOLD+JSEEX+R153+DJT+SP500),eval=TRUE,rank="AIC")
w=attributes(v)
k=Sys.getenv(w$formulas[1]);
k

#DSY
v=dredge(lm(DSY~J203+J200+J835+J211+J210+GOLD+JSEEX+R153+DJT+SP500),eval=TRUE,rank="AIC")
w=attributes(v)
k=Sys.getenv(w$formulas[1]);
k

#FSR
v=dredge(lm(FSR~J203+J200+J835+J211+J210+GOLD+JSEEX+R153+DJT+SP500),eval=TRUE,rank="AIC")
w=attributes(v)
k=Sys.getenv(w$formulas[1]);
k

#GFI
v=dredge(lm(GFI~J203+J200+J835+J211+J210+GOLD+JSEEX+R153+DJT+SP500),eval=TRUE,rank="AIC")
w=attributes(v)
k=Sys.getenv(w$formulas[1]);
k

#GRT
v=dredge(lm(GRT~J203+J200+J835+J211+J210+GOLD+JSEEX+R153+DJT+SP500),eval=TRUE,rank="AIC")
w=attributes(v)
k=Sys.getenv(w$formulas[1]);
k

#HAR
v=dredge(lm(HAR~J203+J200+J835+J211+J210+GOLD+JSEEX+R153+DJT+SP500),eval=TRUE,rank="AIC")
w=attributes(v)

```

```

k=Sys.getenv(w$formulas[1]);
k

#IMP
v=dredge(lm(IMP~J203+J200+J835+J211+J210+GOLD+JSEEX+R153+DJT+SP500),eval=TRUE,rank="AIC")
w=attributes(v)
k=Sys.getenv(w$formulas[1]);
k

#IPL
v=dredge(lm(IPL~J203+J200+J835+J211+J210+GOLD+JSEEX+R153+DJT+SP500),eval=TRUE,rank="AIC")
w=attributes(v)
k=Sys.getenv(w$formulas[1]);
k

#INP
v=dredge(lm(INP~J203+J200+J835+J211+J210+GOLD+JSEEX+R153+DJT+SP500),eval=TRUE,rank="AIC")
w=attributes(v)
k=Sys.getenv(w$formulas[1]);
k

#LBH
v=dredge(lm(LBH~J203+J200+J835+J211+J210+GOLD+JSEEX+R153+DJT+SP500),eval=TRUE,rank="AIC")
w=attributes(v)
k=Sys.getenv(w$formulas[1]);
k

#LBT
v=dredge(lm(LBT~J203+J200+J835+J211+J210+GOLD+JSEEX+R153+DJT+SP500),eval=TRUE,rank="AIC")
w=attributes(v)
k=Sys.getenv(w$formulas[1]);
k

#LON
v=dredge(lm(LON~J203+J200+J835+J211+J210+GOLD+JSEEX+R153+DJT+SP500),eval=TRUE,rank="AIC")
w=attributes(v)
k=Sys.getenv(w$formulas[1]);
k

#MSM
v=dredge(lm(MSM~J203+J200+J835+J211+J210+GOLD+JSEEX+R153+DJT+SP500),eval=TRUE,rank="AIC")
w=attributes(v)
k=Sys.getenv(w$formulas[1]);
k

#MTN
v=dredge(lm(MTN~J203+J200+J835+J211+J210+GOLD+JSEEX+R153+DJT+SP500),eval=TRUE,rank="AIC")
w=attributes(v)
k=Sys.getenv(w$formulas[1]);
k

#MUR
v=dredge(lm(MUR~J203+J200+J835+J211+J210+GOLD+JSEEX+R153+DJT+SP500),eval=TRUE,rank="AIC")
w=attributes(v)
k=Sys.getenv(w$formulas[1]);
k

#NPN
v=dredge(lm(NPN~J203+J200+J835+J211+J210+GOLD+JSEEX+R153+DJT+SP500),eval=TRUE,rank="AIC")
w=attributes(v)
k=Sys.getenv(w$formulas[1]);
k

#NED
v=dredge(lm(NED~J203+J200+J835+J211+J210+GOLD+JSEEX+R153+DJT+SP500),eval=TRUE,rank="AIC")
w=attributes(v)
k=Sys.getenv(w$formulas[1]);
k

#NTC

```

```

v=dredge(lm(NTC~J203+J200+J835+J211+J210+GOLD+JSEEX+R153+DJT+SP500),eval=TRUE,rank="AIC")
w=attributes(v)
k=Sys.getenv(w$formulas[1]);
k

#OML
v=dredge(lm(OML~J203+J200+J835+J211+J210+GOLD+JSEEX+R153+DJT+SP500),eval=TRUE,rank="AIC")
w=attributes(v)
k=Sys.getenv(w$formulas[1]);
k

#PIK
v=dredge(lm(PIK~J203+J200+J835+J211+J210+GOLD+JSEEX+R153+DJT+SP500),eval=TRUE,rank="AIC")
w=attributes(v)
k=Sys.getenv(w$formulas[1]);
k

#PPC
v=dredge(lm(PPC~J203+J200+J835+J211+J210+GOLD+JSEEX+R153+DJT+SP500),eval=TRUE,rank="AIC")
w=attributes(v)
k=Sys.getenv(w$formulas[1]);
k

#REM
v=dredge(lm(REM~J203+J200+J835+J211+J210+GOLD+JSEEX+R153+DJT+SP500),eval=TRUE,rank="AIC")
w=attributes(v)
k=Sys.getenv(w$formulas[1]);
k

#RMH
v=dredge(lm(RMH~J203+J200+J835+J211+J210+GOLD+JSEEX+R153+DJT+SP500),eval=TRUE,rank="AIC")
w=attributes(v)
k=Sys.getenv(w$formulas[1]);
k

#SAB
v=dredge(lm(SAB~J203+J200+J835+J211+J210+GOLD+JSEEX+R153+DJT+SP500),eval=TRUE,rank="AIC")
w=attributes(v)
k=Sys.getenv(w$formulas[1]);
k

#SLM
v=dredge(lm(SLM~J203+J200+J835+J211+J210+GOLD+JSEEX+R153+DJT+SP500),eval=TRUE,rank="AIC")
w=attributes(v)
k=Sys.getenv(w$formulas[1]);
k

#SAP
v=dredge(lm(SAP~J203+J200+J835+J211+J210+GOLD+JSEEX+R153+DJT+SP500),eval=TRUE,rank="AIC")
w=attributes(v)
k=Sys.getenv(w$formulas[1]);
k

#SOL
v=dredge(lm(SOL~J203+J200+J835+J211+J210+GOLD+JSEEX+R153+DJT+SP500),eval=TRUE,rank="AIC")
w=attributes(v)
k=Sys.getenv(w$formulas[1]);
k

#SHP
v=dredge(lm(SHP~J203+J200+J835+J211+J210+GOLD+JSEEX+R153+DJT+SP500),eval=TRUE,rank="AIC")
w=attributes(v)
k=Sys.getenv(w$formulas[1]);
k

#SBK
v=dredge(lm(SBK~J203+J200+J835+J211+J210+GOLD+JSEEX+R153+DJT+SP500),eval=TRUE,rank="AIC")
w=attributes(v)
k=Sys.getenv(w$formulas[1]);
k

```

```

#SHF
v=dredge(lm(SHF~J203+J200+J835+J211+J210+GOLD+JSEEX+R153+DJT+SP500),eval=TRUE,rank="AIC")
w=attributes(v)
k=Sys.getenv(w$formulas[1]);
k

#BVT
v=dredge(lm(BVT~J203+J200+J835+J211+J210+GOLD+JSEEX+R153+DJT+SP500),eval=TRUE,rank="AIC")
w=attributes(v)
k=Sys.getenv(w$formulas[1]);
k

#TBS
v=dredge(lm(TBS~J203+J200+J835+J211+J210+GOLD+JSEEX+R153+DJT+SP500),eval=TRUE,rank="AIC")
w=attributes(v)
k=Sys.getenv(w$formulas[1]);
k

#TRU
v=dredge(lm(TRU~J203+J200+J835+J211+J210+GOLD+JSEEX+R153+DJT+SP500),eval=TRUE,rank="AIC")
w=attributes(v)
k=Sys.getenv(w$formulas[1]);
k

#####
summary(lm(StockData[1:n,1] ~ DJT + J200 + J203 + J835));
summary(lm(StockData[1:n,2] ~ GOLD + J835 + R153));
summary(lm(StockData[1:n,3] ~ DJT + J200 + J203 + JSEEX + R153));
summary(lm(StockData[1:n,4] ~ GOLD + J210 + SP500));
summary(lm(StockData[1:n,5] ~ DJT + GOLD + J210 + SP500));
summary(lm(StockData[1:n,6] ~ GOLD + J200 + J203 + J210 + J835));
summary(lm(StockData[1:n,7] ~ DJT + J210 + JSEEX + SP500));
summary(lm(StockData[1:n,8] ~ GOLD + J203 + J210 + J211 + J835 + SP500));
summary(lm(StockData[1:n,9] ~ J200 + J203 + J835));
summary(lm(StockData[1:n,10] ~ DJT + J200 + J203 + J835));
summary(lm(StockData[1:n,11] ~ J200 + SP500));
summary(lm(StockData[1:n,12] ~ GOLD + J211 + JSEEX + R153 + SP500));
summary(lm(StockData[1:n,13] ~ J203 + J210 + J835 + JSEEX + SP500));
summary(lm(StockData[1:n,14] ~ J200 + J210 + J211 + J835));
summary(lm(StockData[1:n,15] ~ GOLD + J210 + JSEEX + SP500));
summary(lm(StockData[1:n,16] ~ J200 + J203 + J835 + R153));
summary(lm(StockData[1:n,17] ~ GOLD + J210 + JSEEX + SP500));
summary(lm(StockData[1:n,18] ~ GOLD + J200 + J211 + SP500));
summary(lm(StockData[1:n,19] ~ GOLD + J200 + J203 + JSEEX + R153));
summary(lm(StockData[1:n,20] ~ DJT + J200 + J210 + R153 + SP500));
summary(lm(StockData[1:n,21] ~ J203 + J210 + J211 + J835 + SP500));
summary(lm(StockData[1:n,22] ~ J835 + JSEEX + R153 + SP500));
summary(lm(StockData[1:n,23] ~ J200 + J203 + J211 + J835 + JSEEX + R153 + SP500));
summary(lm(StockData[1:n,24] ~ J200 + J203 + J211));
summary(lm(StockData[1:n,25] ~ DJT + J200 + J203 + J211 + JSEEX + R153 + SP500));
summary(lm(StockData[1:n,26] ~ J200 + J203 + J210 + J835 + R153 + SP500));
summary(lm(StockData[1:n,27] ~ DJT + GOLD + J211 + JSEEX + R153 + SP500));
summary(lm(StockData[1:n,28] ~ J203 + J210 + J835));
summary(lm(StockData[1:n,29] ~ J200 + J203 + JSEEX));
summary(lm(StockData[1:n,30] ~ J203 + J210 + J211 + SP500));
summary(lm(StockData[1:n,31] ~ J203 + J210 + JSEEX + SP500));
summary(lm(StockData[1:n,32] ~ GOLD + J200 + J203 + JSEEX + R153));
summary(lm(StockData[1:n,33] ~ GOLD + J203 + J210 + J211 + SP500));
summary(lm(StockData[1:n,34] ~ J835));
summary(lm(StockData[1:n,35] ~ J211 + JSEEX + R153));
summary(lm(StockData[1:n,36] ~ DJT + J200 + J210 + J211 + R153));
summary(lm(StockData[1:n,37] ~ DJT + J200 + J203 + J835 + R153));
summary(lm(StockData[1:n,38] ~ J210 + R153));
summary(lm(StockData[1:n,39] ~ J200 + J203 + J211 + JSEEX + SP500));
summary(lm(StockData[1:n,40] ~ GOLD + J200 + J203 + J835 + SP500));
summary(lm(StockData[1:n,41] ~ J200 + J203 + J211 + JSEEX + R153));
summary(lm(StockData[1:n,42] ~ GOLD + J200 + J203 + J210));
summary(lm(StockData[1:n,43] ~ J211 + JSEEX + SP500));

```

```
summary(lm(StockData[1:n,44] ~ J200 + J203 + J210 + R153));
```

The code is repeated for BIC

ICOMP ALL POSSIBLE REGRESSION CODE

```
library(MuMIn)
library(icompl)

#ABL
v=dredge(lm(ABL~J203+J200+J835+J211+J210+GOLD+JSEEX+R153+DJT+SP500),eval=TRUE,rank="ICOMP.lm")
w=attributes(v)
k=Sys.getenv(w$formulas[1]);
k
```

The code is repeated for all the other stocks

PORTFOLIOS' PDI AND ICOMP CODE

```
#Portfolio 1 Across sector market weight
MSM1=0.03*MSM
APN1=0.06*APN
CFR1=0.24*CFR
TBS1=0.06*TBS
SHF1=0.06*SHF
NTC1=0.04*NTC
SAB1=0.07*SAB
BVT1=0.08*BVT
SOL1=0.36*SOL
x=c(MSM1,APN1,CFR1,TBS1,SHF1,NTC1,SAB1,BVT1,SOL1)
Portfolio1=matrix(x,nrow=103,ncol=9)
princomp(Portfolio1)
plot(princomp(Portfolio1))
d1<-eigen(cov(Portfolio1))
plot(d1$values,type="h")
x1=seq(1,9)
PDI1=sum(d1$values/sum(d1$values)*x1)*2-1
PDI1
ICOMP1=(ncol(Portfolio1)/2*log(sum(diag(cov(Portfolio1))/ncol(Portfolio1)))) - 1/2*log(det(cov(Portfolio1)))
ICOMP1

#Portfolio 2 Across sectors
MSM2=0.11*MSM
APN2=0.11*APN
CFR2=0.11*CFR
TBS2=0.11*TBS
SHF2=0.11*SHF
NTC2=0.11*NTC
SAB2=0.11*SAB
BVT2=0.11*BVT
SOL2=0.11*SOL
x=c(MSM2,APN2,CFR2,TBS2,SHF2,NTC2,SAB2,BVT2,SOL2)
Portfolio2=matrix(x,nrow=103,ncol=9)
princomp(Portfolio2)
plot(princomp(Portfolio2))
d2<-eigen(cov(Portfolio2))
plot(d2$values,type="h")
x2=seq(1,9)
PDI2=sum(d2$values/sum(d2$values)*x2)*2-1
PDI2
ICOMP2=(ncol(Portfolio2)/2*log(sum(diag(cov(Portfolio2))/ncol(Portfolio2)))) - 1/2*log(det(cov(Portfolio2)))
ICOMP2
```

```

#Portfolio 3 Financials
ABL3=0.143*ABL
ASA3=0.143*ASA
FSR3=0.143*FSR
INP3=0.143*INP
NED3=0.143*NED
RMH3=0.143*RMH
SBK3=0.143*SBK
x=c(ABL3,ASA3,FSR3,INP3,NED3,RMH3,SBK3)
Portfolio3=matrix(x,nrow=103,ncol=7)
princomp(Portfolio3)
plot(princomp(Portfolio3))
d3<-eigen(cov(Portfolio3))
plot(d3$values,type="h")
x3=seq(1,7)
PDI3=sum(d3$values/sum(d3$values)*x3)*2-1
PDI3
ICOMP3=(ncol(Portfolio3)/2*log(sum(diag(cov(Portfolio3))/ncol(Portfolio3)))) - 1/2*log(det(cov(Portfolio3)))
ICOMP3

#Portfolio 4 Resources equal weight
AGL4=0.11*AGL
AMS4=0.11*AMS
ARI4=0.11*ARI
ANG4=0.11*ANG
BIL4=0.11*BIL
GFI4=0.11*GFI
HAR4=0.11*HAR
IMP4=0.11*IMP
LON4=0.11*LON
x=c(AGL4,AMS4,ARI4,ANG4,BIL4,GFI4,HAR4,IMP4,LON4)
Portfolio4=matrix(x,nrow=103,ncol=9)
princomp(Portfolio4)
plot(princomp(Portfolio4))
d4<-eigen(cov(Portfolio4))
plot(d4$values,type="h")
x4=seq(1,9)
PDI4=sum(d4$values/sum(d4$values)*x4)*2-1
PDI4
ICOMP4=(ncol(Portfolio4)/2*log(sum(diag(cov(Portfolio4))/ncol(Portfolio4)))) - 1/2*log(det(cov(Portfolio4)))
ICOMP4

#Portfolio 5 Resources market weight
AGL5=0.28*AGL
AMS5=0.12*AMS
ARI5=0.02*ARI
ANG5=0.07*ANG
BIL5=0.33*BIL
GFI5=0.05*GFI
HAR5=0.02*HAR
IMP5=0.08*IMP
LON5=0.03*LON
x=c(AGL5,AMS5,ARI5,ANG5,BIL5,GFI5,HAR5,IMP5,LON5)
Portfolio5=matrix(x,nrow=103,ncol=9)
princomp(Portfolio5)
plot(princomp(Portfolio5))
d5<-eigen(cov(Portfolio5))
plot(d5$values,type="h")
x5=seq(1,9)
PDI5=sum(d5$values/sum(d5$values)*x5)*2-1
PDI5
ICOMP5=(ncol(Portfolio5)/2*log(sum(diag(cov(Portfolio5))/ncol(Portfolio5)))) - 1/2*log(det(cov(Portfolio5)))
ICOMP5

#Portfolio 6 Financial Market weight
ABL6=0.07*ABL
ASA6=0.25*ASA
FSR6=0.28*FSR

```

```

INP6=0.07*INP
NED6=0.19*NED
RMH6=0.1*RMH
SBK6=0.04*SBK
x=c(ABL6,ASA6,FSR6,INP6,NED6,RMH6,SBK6)
Portfolio6=matrix(x,nrow=103,ncol=7)
princomp(Portfolio6)
plot(princomp(Portfolio6))
d6<-eigen(cov(Portfolio6))
plot(d6$values,type="h")
x6=seq(1,7)
PDI6=sum(d6$values/sum(d6$values)*x6)*2-1
PDI6
ICOMP6=(ncol(Portfolio6)/2*log(sum(diag(cov(Portfolio6))/ncol(Portfolio6)))) - 1/2*log(det(cov(Portfolio6)))
ICOMP6

```