Diversification measures and the optimal number of Stocks in a portfolio: An information theoretic explanation

Adeola Oyenubi

ERSA working paper 666

February 2017

Economic Research Southern Africa (ERSA) is a research programme funded by the National Treasury of South Africa.

Diversification measures and the optimal number of Stocks in a portfolio: An information theoretic explanation

Adeola Oyenubi*

February 16, 2017

Abstract

This paper provides a plausible explanation for why the optimum number of stocks in a portfolio is elusive, and suggests a way to determine this optimal number. Diversification is dependent on the number of stocks in a portfolio and the correlation structure. Adding stocks to a portfolio increases the level of diversification, and consequently leads to risk reduction. However the risk reduction effect dissipates after a certain number of stocks, beyond which additional stocks do not contribute to risk reduction. To explain this phenomenon, this paper investigates the relationship between portfolio diversification and concentration using a genetic algorithm. To quantify diversification, we use the Portfolio Diversification Index (PDI). In the case of concentration, we introduce a new quantification method. Concentration is quantified as complexity of the correlation matrix[1]. The proposed method quantifies the level of dependency (or redundancy) between stocks in a portfolio. By contrasting the two methods it is shown that the optimal number of stocks that optimizes diversification depends on both number of stocks and average correlation. Our result shows that, for a given universe, there is a set of Pareto optimal portfolios each containing a different number of stocks that simultaneously maximizes diversification and minimizes concentration. The preferred portfolio among the Pareto set will depend on the preference of the investor. Our result also suggests that an ideal condition for the optimal number of stocks is when the variance reduction as a result of adding a stock is off-set by the the variance contribution of complexity.

Keywords: Information Theory; Diversification; Genetic Algorithm; Portfolio optimization; Principal Component Analysis; Simulation methods; Maximum Diversification Index.

*School of Economics, University of Cape Town. Email: OYNADE001@myuct.ac.za

1 Introduction

Achieving the full benefit of diversification (namely, variance reduction) in a portfolio of securities is desirable in portfolio management. The literature suggests that increasing the number of stocks in stock portfolios is beneficial, in terms of diversification. However, there is no consensus as to what the optimal number of stocks should be. Evans & Archer (1968) put this number between 10 and 12 stocks. Meir (1987) challenges this belief, and finds that between 30 and 40 stocks are needed to achieve full diversification. In Tang (2004)'s review of the literature, the recommendation is between 10 and 40 stocks. The author suggests that a portfolio size of 20 is needed to eliminate 95% of the diversifiable risk. The common theme among all these studies is that a stock portfolio is made up of diversifiable and non-diversifiable risks. Adding stocks to a portfolio helps deal with diversifiable risks, but once all diversifiable risks have been accounted for, additional stocks hold no benefit, in terms of diversification.

While this paper agrees that there is no magic number that achieves full diversification in every situation (or under every market condition), our main contribution is to provide an alternative explanation as to why this is the case. To achieve this, we introduce a novel methodology to quantify diversification or concentration. By contrasting the proposed measure with a known method of quantifying diversification, an alternative explanation is provided as to why the diversification benefit depends not only on the number of stocks, but also the correlation structure.

Our approach involves comparing a measure of diversification with a measure that captures lack of diversification (or variation redundancy). The trade-off between these two factors, as stocks are added to a portfolio, suggests the optimal number of stocks that exhaust the benefits of diversification. Our main conjecture is that, on the one hand, as stocks are added to a portfolio this increases its level of diversification by introducing independent sources of variation to the portfolio, which eliminates diversifiable risk. On the other hand, this also increases the level of redundancy in the portfolio, by amplifying the variation that is common to all stocks, or the non-diversifiable risk. This component is captured by our proposed measure, which we call portfolio complexity. The net effect is such that, beyond a certain number of stocks, the marginal diversification benefit of the last stock is off-set by its complexity contribution (or variation redundancy), so that adding more stocks beyond this point holds no diversification benefit.

2 Quantifying diversification

Diversification is one of the most important tenets of portfolio theory. However, finding a metric for measuring the degree of diversification of a given portfolio has been elusive. Since the pioneering work of Markowitz (1952), different methods have been proposed, in an attempt to measure diversification across portfolios. Despite these proposals, there has not been consensus on a metric

to measure diversification (Frahm & Wiechers 2011; Rudin & Morgan 2006). The lack of consensus is not because of conflicting definitions of the concept, rather it arises out of the problem of how to measure the quantity defined. A portfolio is said to be diversified if its sources of variation are independent. It is important to note here that sources of variation for stock portfolios can be thought of in two ways. Some diversification measures assume that individual stocks constitute sources of variation, while the more recent measures think of sources of variation in terms of factors. These factors are orthogonal linear combinations of stock variations.

The relationship between the degree of diversification, and variance, is what makes diversification an important concept. According to Frahm & Wiechers (2011), "It is the diversification effect among different assets that seems to contribute to portfolio performance". One interpretation of this is that the more independent sources of variation there are in a portfolio, the less likely it will be heavily vulnerable to individual component shocks. This in turn keeps the volatility as low as possible.

3 Diversification measures

A number of diversification measures have been introduced in the literature. Broadly speaking, these measures can be classified into two categories. The first category measures diversification relative to some market index, by the use of factor analysis, while the second category measures it independent of an index.

The second category is desirable because there are cases where the index portfolio itself is not well diversified. An example is the case of developing markets, where a particular sector of the market might be responsible for a disproportionately large percentage of the total variation in the market (Van Heerden, *et. al.* 2008). The category that measures diversification independent of the market portfolio can be further divided into three sub-categories. The first category assumes independence among the primary constituents of a portfolio (stocks), and then defines a diversified portfolio as a portfolio with a weight that is uniformly distributed over its sources of variation (e.g. Herfindal index (Hovenkamp 1985)), and related measures based on portfolio weights. These measures are also referred to as concentration measures (Divecha, *et al.* 1992).

The second sub-category ignores portfolio weights, instead, its quantification is based on the covariance/correlation of stocks in the portfolio. This sub-category uses some metric to define independence among unique variation factors in a portfolio, e.g. Principal Component based methods like the Portfolio Diversification Index (PDI) (Rudin & Morgan 2006). The third sub-category combines both approaches, i.e. it first defines independent sources of variation in the portfolio using factor analysis, and then distributes weights such that the portfolio's risk is not dependent on a single source of variation. An example is the method introduced by Meucci (2010) which uses the principal component analysis approach with information theory in quantifying diversification. The

merits and demerits of these approaches are discussed in Frahm & Wiechers (2011). In this paper we make use of the PDI measure to capture the level of diversification in a portfolio. The choice of PDI is based on the fact that, like our proposed complexity measure, also falls into the category of diversification quantifiers that ignore portfolio weights[2]. The subsequent analysis is therefore based on equally weighted portfolios.

3.1 Portfolio Diversification Index (PDI)

PDI is constructed using principal component analysis. It is designed to evaluate the effective number of independent variation components in a portfolio. Under this method, the principal components of a portfolio's return is calculated and a weight is attached to each principal component in the following way;

$$PDI = 2\sum_{k=1}^{n} kW_i - 1 \ldots\ldots\ldots \tag{1}$$

$$W_i = \frac{\lambda_i}{\sum \lambda_i}$$

where n is the number of principal components in a portfolio and λ_i are the ordered and normalized covariance or correlation eigenvalues. This approach expresses a portfolio as a linear combination of variation factors or principal portfolios that are by definition orthogonal (Meussi 2009). The eigenvalues are the variances of these principal portfolios. Therefore, W_i is the fraction of the original portfolio's total variance that is attributable to the i^{th} principal portfolio.

PDI therefore measures diversification as the centre of mass of principal components or relative strength vector. It measures how front loaded the vector $\{\lambda_i\}i = 1 \ldots\ldots n$ is. It is also bounded in the interval $1 \leq PDI \leq n$ so that PDI can be interpreted as the number of independent sources of variation in the portfolio. When all the eigenvalues (λ_i) in a portfolio are equal (e.g. if the covariance matrix is an identity matrix) then the $PDI = n$. This means that there are as many orthogonal factors of variation as there are stocks in the portfolio. Essentially the bigger the PDI value the more independent sources of variation there are in the portfolio, and consequently the more diversified the portfolio is. For details of implementation of this index see Rudin & Morgan (2006).

3.2 Complexity measure

We define an alternative measure for quantifying diversification. Unlike concentration measures that are based on weights (*Herfindal index*), our proposed measure expresses concentration as a function of the dependency structure of

[2]It is worth mentioning that Kirchner and Zunckel (2011) claim portfolio weight can be accommodated under PDI by considering weighted returns.

returns in the portfolio. This measure simply quantifies lack of diversification as the complexity of the variance-covariance or correlation matrix of returns.

The complexity of a random vector is a measure of the interaction or dependency between its components. For a multivariate normal distribution of dimension n with joint probability density function (pdf) $f(x) = f(x_1, x_2, x_3, \ldots\ldots x_n)$ and marginal pdfs $f_j(x_j), j = 1, 2, 3, \ldots\ldots, n$. Information measure of dependence between the random variables $x_1, x_2, x_3, \ldots\ldots x_n$ is given by;

$$I(x) = I(x_1, x_2, x_3, \ldots\ldots x_n) = \tag{2}$$
$$E_f[\log \frac{f(x_1, x_2, x_3, \ldots\ldots x_n)}{f(x_1), f(x_2,), f(x_3,) \ldots\ldots, f(x_n)}]$$

where $I(x)$ is the Kullback-Leibler (KL) (1951) information divergence against independence. $I(x)$ is a measure of expected dependency among the component variables, and has the following properties;

- $I(x) \geq 0$ .i.e the expected dependency is non-negative
- $f(x_1, x_2, x_3, \ldots\ldots x_n) = f(x_1), f(x_2,), f(x_3,) \ldots\ldots, f(x_n)$, if and only if the variables are statistically independent, in which case the quotient in equation (2) will be 1 and the log of it will be zero.

Assuming the n-variate random variables are equity returns (or any other security), it means that $I(x)$ is the measure of expected dependency between stocks in the portfolio. Hence, $I(x)$ for a portfolio of stocks is zero if all the stocks are statistically independent. However, for stock portfolios we expect $I(x)$ to always be greater than zero because of the theory of factor models. For a given portfolio or market, higher values of $I(x)$ correspond to higher levels of dependency.

KL divergence in equation 1 is related to Shanon (1948) entropy by the identity;

$$I(x) = \sum_{j=1}^{n} H(x_j) - H(x_1, x_2, x_3, \ldots\ldots x_n) \tag{3}$$

where $H(x_j)$ is the marginal entropy and $H(x_1, x_2, x_3, \ldots\ldots x_n)$ represents the joint entropy. Van Edem (1971) provides the definition for information complexity of a covariance matrix Σ of normally distributed random variables.

$$I(x) = \sum_{j=1}^{n} \left[\frac{1}{2}\log(2\pi) + \frac{1}{2}\log(\sigma_j^2) + \frac{1}{2}\right] - \left[\frac{1}{2}\log(2\pi) + \frac{1}{2}\log|\Sigma| + \frac{1}{2}\right] \tag{4}$$

This reduces to;

$$C_0(\Sigma) = \frac{1}{2}\sum_{j=1}^{n}\log(\sigma_j^2) - \frac{1}{2}\log$$

See Bozdogan (2004) for details of the proof. σ_j^2 is the j-th diagonal element of Σ and n is the dimension of Σ.$C_0(\Sigma)$ is zero when Σ is a diagonal matrix and $C_0(\Sigma)$ is infinity if any one of the variables may be expressed as a linear combination of the others. Van Edem (1971) points out that $C_0(\Sigma)$ is not an effective measure of the amount of complexity in Σ since it is not invariant to orthogonal transformations.

The maximal covariance complexity corrects this shortfall. The maximal covariance complexity is the maximum of $C_0(\Sigma)$ under orthonormal transformations of the co-ordinate system (Bozdogan 2004).

$$C_1(\Sigma) = \max_T C_0(\Sigma) = \max_T \{\sum_{j=1}^{n} H(x_j) - H(x_1, x_2, x_3, \ldots\ldots x_n)\} \tag{5}$$

$$= \frac{n}{2}\log\left[\frac{tr(\Sigma)}{n}\right] - \frac{1}{2}\log|\Sigma| \tag{6}$$

where the maximum is taken over the orthonormal transformation T of the overall co-ordinate systems $x_1, x_2, x_3, \ldots\ldots x_n$. $C_1(\Sigma)$ is the upper bound of $C_0(\Sigma)$, it measures both inequality among the variances and the contribution of the covariances of Σ (Van Edem, 1971). $C_1(\Sigma)$ is invariant with respect to scalar multiplication and orthonormal transformation. It is also a monotonically increasing function of n, the dimension of Σ (Magnus and Neudecker 1999). According to Bozdogan (2000) $C_1(\Sigma)$ can be written as;

$$C_1(\Sigma) = \frac{n}{2}\log(\frac{\lambda_a}{\lambda_g}) \tag{7}$$

where λ_a and λ_g are the arithmetic and geometric average of the eigenvalues of Σ. The minimum of $C_1(\Sigma)$ corresponds to the least complex portfolio i.e. $C_1(\Sigma)$ approaches 0 as Σ approaches the identity matrix[3]. On the other hand, high values of $C_1(\Sigma)$ indicate high dependency between the securities in the porfolio. As the level of dependency increases $C_1(\Sigma)$ approaches infinity. In this paper $C_1(\Sigma)$ is referred to as the Information Complexity measure (ICOMP or complexity) which captures the level of redundancy, variation replication, or the concentration of a stock portfolio.

A correlation matrix can also be used to describe complexity (Bozdogan 2004). In this case equation 6 reduces to;

$$C_1(R) \equiv C_0(R) = \frac{n}{2}\log\left[\frac{tr(R)}{n}\right] - \frac{1}{2}\log|R| \tag{8}$$
$$= -\frac{1}{2}\log|R|$$

where R is the correlation matrix of $x_1, x_2, x_3, \ldots\ldots x_n$.

[3] If the covariance matrix of a stock portfolio is an identity matrix then all covariances are zero, meaning that the stock returns are statistically independent and all the variances have the same magnitude. Another possibility is when portfolio contain only one stock.

4 Complexity (ICOMP) as a portfolio concentration measure

Originally, concentration of a portfolio referred to the extent to which distribution of portfolio weights depart from uniform distribution. Concentration indices are known to have a direct relationship with variance of stock portfolios (Oyenubi, 2010). There are a number of concentration measures, but the most widely used is the Herfindahi-Hirschman Index (HHI). Portfolio concentration is related to its complexity, in that the concentration index, as defined by HHI, takes the correlation structure between stocks as a given, and quantifies the departure of the distribution of weights (attached to each stock) from the uniform distribution. In this approach only the weights matter. The HHI index is given by;

$$HHI = \sum_{i=1}^{n} W_i^2 \tag{9}$$

where W_i represents the weight attached to stock i. Note that HHI is minimized when all stocks have the same weight irrespective of their correlation. In terms of quantifying diversification, the concentration index as measured by HHI could be misleading as it implicitly assumes stocks by themselves constitute independent sources of variation. While this assumption makes sense for its application in the industrial organization literature (from where this measure was adopted see Woerheide & Persson (1993)[4]), it makes less sense in the case of stock market analysis where the constituents are more likely to be correlated.

Complexity on the other hand first uses principal component analysis to fish out the truly independent sources of variation and then measures the departure of these variation factors from uniform distribution (equation 7)[5]. This approach is more applicable to security analysis. Complexity can therefore be thought of as a concentration measure that does not ignore the correlation structure of its constituents. The main difference is that while HHI measures dispersion of weights from the uniform distribution complexity measures dispersion of orthogonal factors from uniform distribution[6]

Van Heerden, *et. al.* (2008) notes that portfolio risk is a function of concentration (weighting structure) and covariance between the assets in a portfolio. The authors note that both concentration, as measured by HHI, and covariance, as measured by the correlation matrix, have a positive relationship with variance (the FTSE/JSE All Share Index was examined). While they assess covariance or the dependency structure by a correlation matrix, complexity provides a way of assessing the dependency structure with a single value that summarizes the dependency structure. HHI and variance are expected to be positively related.

[4] Woerheide & Persson (1993) use a variant of the HHI as a diversification measure i.e. 1-HHI

[5] i.e. ratio of arithmetic and geometric mean equals 1 if and only if all values are equal

[6] The suggestion of Kirchner and Zunckel (2011) to incorporate weights into PDI can be accommodated by ICOMP so that we have a concentration measure that accounts for both correlation structure and weights.

In subsequent sections we first investigate the relationship between complexity as a concentration measure and variance. We then examine the relationship between complexity, PDI, concentration, number of stocks, and variance of a portfolio of stocks.

5 Logical Implications of the Relationship between ICOMP, Variance and PDI

The efficacy of any diversification measure is based on its ability to produce portfolios that have lower volatility. In this section we investigate the relationship between complexity, PDI, HHI, and variance. We start by analysing relationships between these quantities that have been established in the literature. For simplicity, we restrict our argument to cases where portfolios are equally weighted.

Rudin & Morgan (2006) show that there is an inverse relationship between PDI and variance. The authors also show that PDI increases with number of stocks, but the marginal effect decreases as more stocks are added to the portfolio (see also Diyarbakirlioglu & Satman (2013)). Variance increases with concentration (measured by HHI) (Van Heerden, *et. al.* 2008; Oyenubi 2010) and concentration decreases with number of stocks.

Going by the definition of complexity (being a measure of concentration), one would expect a negative relationship between PDI and complexity. This is because while PDI counts the number of independent factors in a portfolio, complexity assigns a number to the degree of dependency among portfolio constituents. Intuitively the expectation is that a portfolio with a high PDI will have a low complexity value and vice versa. That is, as the number of independent factors driving variation in a portfolio increases, the degree of dependency among its primary constituents (stocks in this case) decreases. Another way to express this is to note that, as complexity increases, the number of truly independent sources of variation in the portfolio falls, resulting in lower PDI values. Therefore complexity should be inversely proportional to PDI.

There is also a key similarity between complexity and PDI theoretically. Both measures quantify diversification by using the variance (eigenvalues) of the principal components of a portfolio. When the eigenvalues corresponding to all principal components are equal, the proportion of variation attributable to each orthogonal source is the same i.e. the distribution of eigenvalues is uniform. This represents a well-diversified portfolio. Departure from this uniform distribution of orthogonal variation sources represents lack of diversification (i.e. it signals disproportionate exposure to one or some orthogonal factor relative to the others).

PDI measures how uniform eigenvalues are, by determining how front loaded the eigenvalues are. Specifically W_i in equation 1 describes the *pdf* of eigenvalues. If these are uniformly distributed then $W_1 = W_2 = W_3 = \cdots = W_n$ and the portfolio attains the highest rank possible (i.e. $PDI = n$ the number of stocks

in the portfolio). Departure from uniformly distributed eigenvalues results in lower ranking.

Complexity on the other hand measures how uniform eigenvalues are by comparing the arithmetic and geometric mean of eigenvalues in equation 7. The inequality between arithmetic and geometric mean is such that for any list of n non-negative real numbers say $\lambda_1, \lambda_2, \lambda_3 \ldots \lambda_n$

$$\frac{\lambda_1 + \lambda_2 + \lambda_3 + \cdots \lambda_n}{n} \geq \sqrt[n]{\lambda_1 * \lambda_2 * \lambda_3 * \ldots * \lambda_n}$$

This means that the arithmetic mean is always greater than or equal to the geometric mean and equality holds if and only if $\lambda_1 = \lambda_2 = \lambda_3 = \cdots = \lambda_n$. Consequently, the term $\frac{\lambda_a}{\lambda_g}$ in equation 7 has a lower bound of 1. This is the case where eigenvalues are uniformly distributed, and this is analogous to the case where normalized eigenvalues are equal (i.e. $W_1 = W_2 = W_3 = \cdots = W_n$). However unlike PDI uniformly distributed eigenvalues obtain the value zero under complexity (the least complex portfolio) and departure from uniformity is given higher rank. This further explains the expected inverse relationship between ICOMP and PDI

Even though HHI and complexity are both concentration measures, the relationship between complexity and variance may not be similar to the relationship between HHI and variance, since the complexity is based on correlation structure and HHI is based on weights. The logic that suggests a positive relationship between complexity and variance ignores the factor that drives complexity. Complexity could be driven by number of stocks, or average correlation.

If the number of stocks is held constant, an increase in pairwise correlation between stocks should increase complexity and, as a result, the variance. However, when comparing complexity for portfolios with different numbers of stocks, the relationship is slightly different. Complexity is a monotonic function of the number of stocks (Magnus and Neudecker 1999). Therefore, complexity will always increase when stocks are added to a portfolio. This increased complexity may not necessarily result in higher variance, as one may expect, because adding stocks to a portfolio reduces variance on average. This is in contrast to concentration as measured by HHI, which decreases with the addition of stocks.

To illustrate the idea that complexity has different effects on variance, depending on whether it is driven by number of stocks or higher correlation, consider the following example: Assume an equally weighted portfolio of 2 stocks with correlation structure given by

$$\begin{pmatrix} 1 & \rho_{21} \\ \rho_{12} & 1 \end{pmatrix} \tag{10}$$

Where $\rho_{12} = \rho_{21}$Adding a third stock that has the same correlation with the other stocks in the portfolio (i.e. $\rho_{12} = \rho_{21} = \rho_{13} = \rho_{23}$), will result in increased complexity[2], because complexity is a monotonic function of the number

[1] This measure was originally introduced in the author's Master's thesis under the supervision of A.E Clark and C.G Troskie (see Oyenubi (2010) for details)

[2] The magnitude of the increase will depend on the correlation value, as we will show later.

of stocks. The implication of this is that variance (which on average reduces with addition of stocks over some range of values) may reduce, while complexity increases. Therefore, in the simplified case of equally weighted portfolios, the relationship between variance and complexity may depend on the number of stocks in the portfolio. Alternatively, instead of adding another stock to the portfolio, assume that the correlation between the 2 stocks can be increased. This increase will result in higher complexity and consequently higher variance. Therefore complexity, when driven by correlation, will have the expected positive relationship with variance. If, however, it is driven by number of stocks, the relationship may be positive or negative.

The fact that complexity penalizes portfolios with a higher number of stocks is logical. The third stock added in the last example will not increase the potential diversification benefit of the portfolio. In fact the addition of that stock is equivalent to doubling the weight attached to a stock already in the portfolio. This addition amplifies the variation already represented in the portfolio, therefore it holds no diversification benefit relative to the initial portfolio. The next section uses simulation to pin down how these quantities vary with each other.

5.1 Artificial Simulation Example

To get a better idea of the relationship between variance and the diversification measures (i.e. PDI and complexity) a simulation study is conducted. First, correlation matrices of dimensions $n = 4$ to 16 are constructed. Then for each n, the pairwise correlation coefficient ρ, which is set to the same value for all pairwise correlation in each matrix is varied so that $\rho = 0.1, 0.2 \ldots\ldots\ldots, 0.9$, creating 9 different correlation structures for each n. For example, the first set of correlation matrices with $n = 4$ are nine 4 by 4 matrices, each matrix having correlation coefficients 0.1 to 0.9 for all pairwise correlations. This is then replicated for $n = 5$ to 16.

Using these correlation matrices, returns are simulated. Each return is distributed $N(0, 1)$.In this set up the correlation matrix of the simulated returns is given by the simulated correlation matrices described. Such set up enables us to investigate the relationship between the variance and the different diversification measures, since both the number of stocks and the correlation matrix vary in this artificial setting. Figures 1, 2, 3 and 4 show the results. These figures show variation in variance, ICOMP, PDI, and HHI as pairwise correlation varies from 0.1 to 0.9 and the number of simulated returns vary from 4 to 16. For example each slice of Figure 1 along the correlation axis represents variation in variance and correlation for a given number of stocks. Each slice along the number of stocks axis has an analogous interpretation with the correlation held constant.

Figure 1 shows how that variance varies when both the correlation and the number of constituents increase. Holding the number of stocks constant at any value, there is a positive relationship between the variance and the (average) correlation. On the other hand variation in variance seems to depend on the particular average correlation value. When (average) correlation is low, adding

stocks reduces the variance sharply, up to a certain number of stocks beyond which additional stocks do not help in reducing the variance in any significant way. While this is a standard result, the important point to note is that at higher correlation values, additional stocks do not necessarily result in lower variance. This result agrees with the notion that diversification benefit is lost during market crises, when correlation are typically higher.

In figure 2 variance is replaced with complexity. When correlation increases, complexity increases, however complexity increases more with high correlation and high number of stocks, in comparison to when stocks are fewer and average correlation is lower. The figure suggests that complexity is more responsive to changes in correlation when there are many stocks in the portfolio. The least complex portfolio is the one with low correlation between its constituents or low number of stocks. Complexity therefore increases with both number of stocks and correlation.

Comparing Figure 2 with Figure 1 shows that higher correlation leads to higher complexity, which means higher variance (along the column in Figures 1 and 2, both complexity and variance increase), especially when there are many stocks. However when complexity increases as a result of additional stocks the relationship between complexity and variance depends on the average correlation. When correlation between stocks is low variance initially decreases as complexity increases[3], after which variance remains unchanged as complexity increases (along the row in Figure 1 and 2). When correlation is high the relationship between variance and complexity does not show a regular pattern in this simulation example. Variance seems to vary randomly with increases in complexity.

Figure 3 shows the variation in PDI when both number of stocks and correlation increase. The PDI result looks like a mirror image of the complexity result in Figure 2. PDI reduces as correlation increases, the drop is faster when there are many stocks, compared with when there are few stocks. The number of stocks increases PDI, although this increase is more pronounced when average correlation is low.

Comparing Figures 1, 2 and 3, a lower variance is more likely when correlation is low and the number of stocks is high enough to take variance to its lower limit. At any other point, lower variance is possible by either increasing the number of stocks, or lowering the correlation. However, portfolio managers can only control the number of stocks (in terms of whether or not a stock gets a positive weight). So to drive diversification, the only tool available is number of stocks. Our results suggest that increasing the number of stocks in order to increase diversification is more effective when correlation is low, and may in fact be meaningless when correlation is high. Figure 4 is included to confirm that HHI is insensitive to the correlation structure and only captures complexity that is driven by weights.

[3] This is difficult to see on the graphs because of the magnitude of complexity compared with the other axis, as noted earlier, complexity is a monotonic function of number of stocks.

6 Empirical Implications: Establishing the relationship between PDI, Variance and Complexity Measures via Simulation Study

The previous section established the expected relationship between ICOMP, PDI and variance, both logically, and using a simulated example. In this section we put the theoretical predictions to a test with real stock data. To achieve this, equally weighted portfolios containing a varying number of stocks (1 to 83) were drawn randomly from a universe of 83 stocks. The data used are weekly returns data from S&P 500, from January 2005 to November 2013 (Details of the stocks in the universe are presented in the appendix). The draws were replicated thirty times.

Figures 5, 6 and 7 show variation in average variance, PDI, and ICOMP when the number of stocks in the portfolio is increased from 1 to 83 over the thirty replications. Note that this is analogous to variation along the rows in Figures 1 to 4. The first two graphs (5 and 6) agree with the results in the literature and findings in the previous section (i.e. variance decreases as the number of stocks increase, and PDI increases as the number of stocks increase). It should be noted that at about the 35 stocks mark, reduction in variance as a result of adding another stock completely disappears on average, as indicated by the green line in Figure 1.

The ICOMP-number of stocks relationship in Figure 7 shows the expected result, since ICOMP is a monotonic function of number of stocks. It should, however, be noted that the increase in ICOMP from 2 to 20 stocks results in lower variance (Figure 5), and subsequent increases do not have much effect on variance. Comparing this result with the results of previous analysis suggests that the pairwise correlation coefficient in this universe of stock should be low. Indeed, the correlation coefficient ranges from 0.25 to 0.49, with an average of 0.46 (correlation will be even lower if the period 2008/2009 is excluded because of the market crises in this period). This analysis, and those in the previous section, show that using the number of stocks as a tool to achieve diversification will be effective (over a given range) when pairwise correlation between stocks are generally low. Our previous analysis also suggests that the same strategy will be counterproductive when correlation is high.

The next set of figures show the relationship between diversification measures and variance as the number of stocks increase. Figures 8, 9 and 10 show variation in average ICOMP as a function of average variance, as the number of stocks increase (8), average PDI as a function of average variance as the number of stocks increase (9), and finally, average PDI as a function of average ICOMP as the number of stocks increase (10).

Figure 8 shows an inverse relationship between ICOMP and variance. This relationship is only possible when low pairwise correlation prevails (see Figure 1 and 2).

In Figure 9, PDI has an inverse relationship with variance. Our previous analysis suggests that this is true irrespective of what drives PDI. PDI reduces

when pairwise correlation increases, and this increases with the number of stocks (Figures 1 and 3).The PDI is therefore negatively related to variance, irrespective of what drives it.

The relationship between ICOMP and PDI is positive, as shown in Figure 10. While this may be counterintuitive, given the definition of the two measures, our previous analysis confirms that this is possible only when both PDI and complexity are driven by increased number of stocks (Figure 2 and 3 rows).

We adjust for the effect of the number of stocks, using two methods to recover the expected negative relationship between PDI and ICOMP. The first method involves looking at changes in PDI and ICOMP, as the number of stocks increase. The other approach holds the number of stocks constant at 30 (randomly selected each time) and compares complexity with PDI by repeating the same simulation done above. The result confirms that when the number of stocks is adjusted for, there is a clear inverse relationship between ICOMP and PDI, as shown in Figure 11 and 12. This is in agreement with the definition of complexity and PDI.

Figure 13 shows the relationship between ICOMP and variance when the number of stocks is held constant at 30. The results show an initial decrease in average variance, which later increases with increases in ICOMP. This relationship is, however, weaker than the previous ones. Nonetheless, the local polynomial regression line shows that a positive relationship exists. The weak relationship is because the differences in the average ICOMP are not substantial, since the number of stocks is held constant at 30.

This result supports the notion that when number of stocks is accounted for, ICOMP increases with variance.

To reiterate, the hypothesis is that when stocks are added to a portfolio it has positive diversification benefits, but at the same time it has negative complexity (ICOMP) or concentration effect.

When a portfolio contains few stocks, the effect of a new stock will be such that $MARGINAL\ PDI > MARGINAL\ ICOMP$[4] since this new stock is likely to be less correlated with the portfolio than the stocks already in it. However when more stocks are added, it is more likely that the effect of a new stock will be such that $MARGINAL\ PDI < MARGINAL\ ICOMP$ since the portfolio itself is more correlated with the market when there are many stocks in it.

Our conjecture is that there is a benefit to adding stocks if the increased diversification as a result of adding a new stock (Marginal PDI) outweighs the increased complexity of adding the same stock (marginal ICOMP). At the point where these two effects cancel each other out, adding more stocks is counterproductive, in terms of diversification. Reaching this point of course depends on the dependency structure or average correlation in the portfolio. We explore what this means for the optimal number of stocks required to exhaust the diversification potential of a stock portfolio.

[4]Note that this marginal is calculated as $\frac{PDI[i+1]-PDI[i]}{1}$ where $i = 1, 2 \ldots\ldots 83$. Therefore they represent change in PDI or ICOMP for a change in the number of stocks on average.

7 Trade-off between Complexity (ICOMP) and Diversification (PDI)

As stated earlier both PDI and ICOMP quantify departure of the distribution of eigenvalues from the uniform distribution. However the rankings given to portfolios under these measures are inversely related. PDI captures diversification while ICOMP captures concentration in terms of variation replication as against concentration in weights. Complexity can be used as a criterion for portfolio optimization just like PDI has been used in the literature (see Crezee & Swinkels (2010); Diyarbakirlioglu & Satman (2014)). Our main interest however is in the optimal number of stocks that exhausts the benefits of diversification.

Our conjecture is that the optimum number of stocks for any universe of stocks is achieved at the point where the marginal complexity effect of adding a stock cancels out the marginal diversification effect of the same stock. Adding stocks beyond this optimum number will only increase complexity (ICOMP) and as a consequence variance of a stock portfolio. Since these measures are not quantified on the same scale direct comparison or comparing marginal effects may not be the best way of narrowing down the optimal number of stocks. Our approach is to try to maximize PDI while minimizing ICOMP. i.e.

$$\max_{1 \leq i \leq N} (PDI - ICOMP) \tag{11}$$

When equation 10 is maximized for a portfolio, such a portfolio will simultaneously maximize PDI while keeping ICOMP at its minimum possible value

We first illustrate this idea with our simulation example in section 5.1. As noted earlier ICOMP surface in Figure 2 looks like a mirror image of PDI surface in Figure 3. Figure 14 shows the difference between the values of PDI and complexity at every correlation value and number of stock in the simulation example.

The surface in Figure 14 describes variation in the number of stocks that exhaust the diversification potential of our universe of 16 artificial stocks. The points on this surface represent the difference between PDI and complexity as shown in Figures 2 and 3. When correlation is low, this difference is maximized when all stocks are included in the portfolio. At higher average correlation, the optimal number decreases. At very high correlation, the portfolio should contain the minimum number of stocks possible, as all stocks represent approximately the same variation.

The problem of maximizing PDI while minimizing ICOMP is clearly a multi-objective optimization problem. Since this problem involves competing objectives one should be interested in a set of Pareto optimal solutions (or portfolios). These portfolios are Pareto optimal in the sense that with respect to the objectives of diversification (PDI) and complexity (ICOMP) they are not dominated by other portfolios in the Pareto set and they dominate any other portfolio that can be formed using our universe of stocks. In the Pareto set these optimal portfolios are such that one cannot increase the PDI of any of them without

simultaneously increasing the ICOMP or complexity and vice versa by adding or removing stocks.

The approach suggested by equation 10 is consistent with the idea of scalarizing a multiobjective optimization problem. This approach involves converting a multi-objective optimization problem into a single objective optimization problem so that the solutions of the single objective optimization problem are Pareto optimal solutions to the multi-objective optimization problem (Hwang & Masud, 2012). Furthermore, it is often required that different Pareto optimal solutions can be reached when different parameters are used in scalarizing the multi-objective problem (Hwang & Masud, 2012).

Equation 10 scalirizes the problem by attaching equal weight to both PDI and ICOMP in finding a portfolio that optimizes diversification benefit. There is however no reason why other weighting structures cannot be used with equation 10. For example putting greater weight on PDI relative to ICOMP might be attractive when average correlation between stocks is low (this is consistent with Figure 14) The converse might be appealing when there is high correlation between stocks. One can therefore think of the single objective version as a utility function where the weights attached to PDI and ICOMP reflect the preferences of the investor in terms of the trade-off between PDI and ICOMP In the analysis that follows we consider equation 10 as a scalarized solution to the multi-objective problem and also perform a multi-objective optimization to get the set of Pareto optimal portfolios for our universe of stocks.

7.1 Genetic Algorithm: Search for the optimum number of stocks

Genetic Algorithm (GA) (Holland, 1975) is a search heuristic that mimics natural selection in order to optimize non-linear or non-differentiable objective function(s) (Diyarbakirlioglu & Satman 2013). GA is a stochastic search algorithm inspired by the basic principle of biological evolution and natural selection (Scrucca, 2012).

In this paper we apply the GA method to the problem of finding the optimal number of stocks that optimizes diversification benefit given a universe of stocks. Specifically we are interested in the number of stocks that maximize diversification (PDI) and simultaneously minimize complexity (ICOMP)

We note that the trade-off in equation 10 can potentially be used to construct a portfolio (i.e. equation 11 can be defined in terms of weights as against number of stocks) similar to the way Diversification ratio is used to construct the Most Diversified Portfolio (Choueifaty *et. al.* 2013), or PDI is used to construct a Maximum Diversification Index (MDI) (Diyarbakirlioglu & Satman 2013). However, in this paper we focus on using this trade-off to explain the optimum number of stocks that exhaust the benefit of diversification.

GAs are implemented by first generating (randomly) a population of solutions to an optimization problem. Each solution is referred to as a chromosome.

The desirability of each chromosome is evaluated by using the fitness function (equation 10 in our case). Chromosomes that perform better, as measured by the

fitness function are selected, mated and mutated to construct a new population of solutions (offspring) that are better than the previous generation of solutions. This is based on the fact that an effective method for creating new models is to combine successful features of two or more existing models (Cooper, 2000; 403). This process is continued until the solution cannot be improved upon. For more detail on GA see Czarnitzki & Doherr (2002), Scrucca (2012), and Maschek (2015).

Our analysis makes use of binary encoding i.e. the population of chromosomes (solutions) are random vectors of zeros and ones of length 83 (total number of stocks in the universe). The ones indicate that the stock in that position in the list of 83 stocks is present in the portfolio while zeros indicate that the stock is not.

A combination of elitism (number of chromosomes that are kept to the next generation) and single point crossover is used to generate a new population. For crossover, two random parents are selected. In addition to this, random genes in a chromosome are selected and mutated, and the number of genes that undergo mutation is controlled by a preselected probability of mutation. To make sure that the fitness function is meaningful, all chromosomes are constrained so that they contain at least 2 stocks[5]. The implementation in R (rbga.bin) has another important parameter called "zeroToOneRatio" this parameter controls the ratio of zeros to ones in the initial randomly generated population.

Diyarbakirlioglu & Satman (2013) extend Rudin and Morgan's (2006) work on PDI by using GA to solve for the Maximum Diversification Index (MDI). Their strategy involves finding securities that maximize the PDI. In their set-up, they consider an investment universe of N assets. An investor is interested in forming a portfolio consisting of a maximum of n_i assets, with $n_i < N$. Therefore the investor is interested in picking n assets, $(n \leq n_i)$ that provide the highest diversification or maximize PDI among the possible subsets of $\binom{n_i}{n}$ portfolios (Diyarbakirlioglu & Satman, 2013). Mathematically

$$\max_{1 \leq i \leq 83} \left\{ 2 \sum_{k=1}^{n} kW_i - 1 \right\} \textit{subject to } n - n_i \leq 0 \tag{12}$$

According to Diyarbakirlioglu & Satman (2013) the objective is to determine the security pool that maximizes the risk diversification benefit while minimizing the number of securities. One feature of this approach (which is also a disadvantage) is that n has to be predetermined.

It is possible to perform a similar exercise for complexity by minimizing ICOMP (MICOMP). The maximization problem then becomes a minimization problem, so that we have:

[5] To estimate the covariance or correlation matrix, the portfolio must contain at least 2 stocks. The single objective Genetic Algorithm is implemented with the "genalg" package in R (Willighagen et.al, 2015) while the multi-objective one is implemented with the "nsga2R" package, also in R. Details of the parameters are supplied in subsequent sections.

$$\min_{1\leq i\leq 83}\left\{-\frac{1}{2}\log|R|\right\} \textit{ subject to } n-n_i\geq 0 \tag{13}$$

Figure 15 shows the graph of MICOMP (minimized complexity) and the average complexity (i.e. the MICOMP graph is superimposed on Figure 6, which shows average complexity for randomly selected portfolios of size n). The GA algorithm is implemented using the rbga.bin command of the "genalg" R package. For this exercise the population size is 1500 and 200 iterations were performed. Elitism is set to 1 and the mutation chance is 20%. While Figure 16 shows an analogous graph for PDI and MPDI (maximized PDI).

The result in Figure 16 agrees with the result presented by (Diyarbakirlioglu & Satman 2013) i.e. the GA algorithm is able to select portfolios that optimize diversification benefits, for a given (pre-selected) number of stocks relative to the average performance of randomly selected portfolios of the same size. Furthermore, our result in Figure 15 shows that the same logic applies for ICOMP.

The GA algorithm selects portfolios that minimize ICOMP for a given (pre-selected) number of stocks, relative to the average performance of randomly selected portfolios of the same size.

7.2 Combining ICOMP and PDI Scalarizing the multi-objective function

The approach of Diyarbakirlioglu & Satman (2013) does not, however, provide a way to select the optimal number of stocks. It merely tells us that portfolios that maximize PDI or minimize ICOMP have better diversification potential, compared to randomly selected portfolios of the same size. We explore a method that combines these two objective functions by scalarizing them and weighting them equally. In this method the GA algorithm is used to select a portfolio that simultaneously maximizes diversification benefits by maximizing PDI and minimizes complexity (or concentration) effects by minimizing ICOMP.

$$\max_{1\leq i\leq 83}\left(\left\{2\sum_{k=1}^{n}kW_i-1\right\}-\{-\frac{1}{2}\log|R|\}\right) \tag{14}$$

The objective function in equation 14 will use the trade-off between PDI and ICOMP over the entire range of plausible values for n and pick the combination of stocks that yields the highest PDI and lowest ICOMP or the portfolio that maximizes the difference between the value of PDI and ICOMP. The larger this difference the more diversified our portfolio is in terms of the two objectives. While there is no reason to directly subtract these two indices since they measure different things, the weighting can be thought of as a way to reflect the preferences of the investor over the two objectives. As noted earlier this approach will result in only one of the many Pareto optimal solutions to this problem and this solution only reflects the preference of an investor that weights PDI and ICOMP

equally. This as mentioned earlier may depend on market conditions[6] Indeed a different weighting structure will lead to different results which will also be Pareto optimal given the preferences expressed by the weighting function.

The parameters of the GA model for the scalarized function are the same as those used for the single objective optimization in the last section. The only exception is that the number of iterations is 800 in this case. The result of optimizing equation 14 shows that a portfolio containing 38 stocks achieves the highest trade-off between ICOMP and PDI for our universe of 83 stocks. The PDI value for this portfolio is approximately 18.2 while the ICOMP value is approximately 8.9. We note that this number is not far from the number of stocks beyond which variance does not reduce in Figure 5 ($\pm$ 35 stocks). We also note that Figures 15 and 16 signal this result. The space between the blue and the red dots in these figures represents the diversification gain of minimizing complexity and maximizing PDI respectively at any given number of stocks. It is clear that maximum benefit is achieved simultaneously around 40 stocks. Specifically, Figure 17 shows the result when MICOMP values (in figure 15) are subtracted from MPDI values (in figure 16).

Even though these are different methods, the logic of the results, and the results themselves, are similar. When there are a few stocks (less than 20 for example), the diversification effect is more than the complexity effect, for portfolios that maximize PDI and minimize ICOMP for any given number of stocks. Thus, the difference between PDI and complexity increases as stocks are added. At the other extreme, when there are many stocks (greater than 40, for example) MICOMP is greater than MPDI. This graphs shows that the optimal point for this universe of stocks is a range that extends from about 33 to 40 stocks.

The result is sensitive to the number of iterations (for example for 200, 400, or 600) the results tend to differ slightly. However, in all cases, the best portfolios are those where the PDI measure is greater than the ICOMP measure by around 9 units, and the portfolio can contain between 30 to 40 stocks[7].

The shape of the graph indicates that for a set of portfolios that optimize both PDI and ICOMP for a decision maker who attaches the same weight to the two objectives, the optimal number of stocks can be roughly estimated using our approach.

It is important to note that the result obtained using this approach, on this or a different universe of stocks, or for a different time period is not a "magic number". The result from this analysis is just one of the many possible plausible Pareto optimal portfolios. By changing the weights attached to each objective one can recover a set of Pareto optimal portfolios, each expressing different trade-offs between the two objectives.

Therefore the preceding result is not a magic number for two main reasons.

[6] As noted earlier, one may want to put more weight on ICOMP when the market is volatile (or when there is high correlation between stocks).

[7] This result is because more than one combination of stocks can achieve a difference of 9 between PDI and ICOMP. Stocks or combination of stocks that are more correlated can be used interchangeable to form portfolios with similar diversification potential

First the optimal number of stocks depends on the correlation structure of the market. Second it depends on the investor's preferred trade-off between PDI and ICOMP. This is conveyed by the weights attached to each objective

One can therefore generalize to obtain a complete set of Pareto optimal portfolios. However instead of trying every conceivable combination of weight a multi-objective approach is a neater way of getting similar results.

7.3 Combining ICOMP and PDI: A multi-objective approach

In this section we use a multi-objective genetic algorithm to get the set of Pareto optimal portfolios that maximize PDI and minimize ICOMP. This approach simultaneously optimizes both objectives and gives a set of solutions that are Pareto optimal with respect to the objectives. That is, moving from one of the solutions to the other is a trade-off that suggests different weighting of the two objectives.

We use the Non-dominated Sorting Genetic Algorithm NSGA II (Deb et.al. 2002) implemented in R package "nsga2R" The main programme uses fast non-dominated sorting, and a crowding distance approach is used to maintain diversity of solutions. The programme also uses tournament selection and binary crossover.

The population size used for this optimization is 1500, tournament size is 2, number of iterations (or generations in the nsga2R package) is 400, and crossover probability and mutation probability are 0.2.

The result (in Figure 18) shows the Pareto optimal solutions in our universe of 83 stocks. As noted earlier one can think of the result as showing the trade-off between PDI and ICOMP that reflects different weighting of the two objectives. Any point on the Pareto front shown in Figure 18 is technically viable, but there may be reasons why a decision maker will prefer a particular point on the Pareto front to others. We can therefore narrow the search space by presenting a subjective but reasonable minimum requirement in terms of trade-off between the two objectives.

One such subjective preference is to insist that the trade-off between ICOMP and PDI should be around one to one. The logic for this is that a trade-off that reflect one unit of PDI for more than one unit of ICOMP is clearly counterproductive in terms of diversification and consequently risk reduction. A trade-off that reflects more than one unit of PDI for one unit of ICOMP can also be argued to be suboptimal in that there is room for improvement. To narrow the search we use this subjective condition because we cannot tell from our analysis how many unit changes in PDI represent a unit change in ICOMP and what the implication of such change is for variance.

To make the analysis clearer we run the programme that produced Figure 18 using a different parameter set and overlay the number of stocks in each portfolio on the Pareto front. The result is presented in Figure 19. The first thing to note about Figure 19 is that as expected the number of stocks increases on average as both PDI and ICOMP increase.

This, coupled with the shape of the graph, confirms that the relationship between average PDI and ICOMP shown in Figure 10 also holds for the portfolios on the Pareto front. Using our subjective condition, a decision maker that has that preference should therefore prefer a subset of the portfolios on the Pareto front where the slope of the tangent to Figure 18 or 19 is 1, for reasons stated earlier. Figure 19 shows that portfolios on the Pareto front that contain roughly 33 to 40 stocks will satisfy this condition.

Just like our previous analysis the multi-objective approach with a reasonable assumption leads to a conclusion that the optimal number of stocks for this universe should be somewhere between 30 and 40.

8 Conclusion

This paper introduces a concentration measure that quantifies complexity of a stock portfolio. Complexity, measured by ICOMP, quantifies lack of diversification of a portfolio. It quantifies the level of dependency among the constituent stocks in a portfolio by comparing the distribution of variation sources to the uniform distribution.

This measure is compared with PDI, which is a diversification measure that quantifies the number of independent sources of variation in a portfolio. It is established that these two measures should be inversely related, if both are driven by correlation. The two quantities can, however, have a positive relationship if they are driven by the number of stocks.

By exploiting this relationship, we show that adding a stock to a portfolio has two effects. First, it can improve the level of diversification of the portfolio. Second, it can also increase its complexity. An optimal number of stocks is reached when a portfolio maximizes diversification while simultaneously minimizing complexity. Our result agrees with the notion that there is no unique number when it comes to the optimal number of stocks needed to achieve full diversification. This number depends on the correlation structure in the universe of stocks under consideration, and the trade-off between diversification and complexity. It is for this reason than one expects the optimum number of stocks needed to achieve full diversification to be different for different markets, such as developing versus developed markets. This is a direct consequence of the different correlation structures in these markets.

References

[1] Bozdogan, Hamparsum. "Akaike's information criterion and recent developments in information complexity." *Journal of mathematical psychology* (Elsevier) 44, no. 1 (2000): 62-91.

[2] —. *Statistical data mining and knowledge discovery.* CRC Press, 2004.

[3] Choueifaty, Yves, Tristan Froidure, and Julien Reynier. "Properties of the most diversified portfolio." *Journal of Investment Strategies* 2, no. 2 (2013): 49-70.

[4] Cooper, Ben. "Modelling research and development: how do firms solve design problems?" *Journal of Evolutionary Economics* (Springer) 10, no. 4 (2000): 395-413.

[5] Cremers, KJ Martijn, and Antti Petajisto. "How active is your fund manager? A new measure that predicts performance." *Review of Financial Studies* (Soc Financial Studies) 22, no. 9 (2009): 3329-3365.

[6] Crezee, Dominiek P, and Laurens AP Swinkels. "High-conviction equity portfolio optimization." *Journal of Risk* 13, no. 2 (2011): 57.

[7] Czarnitzki, Dirk, and Thorsten Doherr. "Genetic algorithms: A tool for optimization in econometrics-basic concept and an example for empirical applications." (ZEW discussion Paper) 2002.

[8] Deb, Kalyanmoy, Amrit Pratap, Sameer Agarwal, and TAMT Meyarivan. "A fast and elitist multiobjective genetic algorithm: NSGA-II." *Evolutionary Computation, IEEE Transactions on* (IEEE) 6, no. 2 (2002): 182-197.

[9] Divecha, Arjun B, Jaime Drach, and Dan Stefek. "Emerging markets: A quantitative perspective." *The journal of portfolio management* (Institutional Investor Journals) 19, no. 1 (1992): 41-50.

[10] Diyarbakriliouglu, Erkin, and Mehmet H Satman. "The Maximum Diversification Index." *Journal of Asset Management* (Nature Publishing Group) 14, no. 6 (2013): 400-409.

[11] Evans, John L, and Stephen H Archer. "Diversification and reduction of dispersion: An empirical analysis." *The Journal of Finance* (Wiley Online Library) 23, no. 5 (1968): 761-767.

[12] Frahm, Gabriel, and Christof Wiechers. "On the diversification of portfolios of risky assets." Tech. rep., Discussion papers in statistics and econometrics, 2011.

[13] Holland, John H. "Adaptation in natural and artificial systems: An introduction with application to biology, control and artificial intelligence." *Ann Arbor, University of Michigan Press*, 1975.

[14] Hovenkamp, Herbert J. "Economics and Federal Antitrust Law." 1985.

[15] Hwang, C-L, and Abu Syed Md Masud. *Multiple objective decision making—methods and applications: A state-of-the-art survey.* Vol. 164. Springer Science \& Business Media, 2012.

[16] Kirchner, Ulrich, and Caroline Zunckel. "Measuring Portfolio Diversification." *arXiv preprint arXiv:1102.4722*, 2011.

[17] Kullback, Solomon, and Richard A Leibler. "On information and sufficiency." *The annals of mathematical statistics* (JSTOR), 1951: 79-86.

[18] Lin, Dan, Xiaoming Li, and Minqiang Li. "A genetic algorithm for solving portfolio optimization problems with transaction costs and minimum transaction lots." In *Advances in Natural Computation*, 808-811. Springer, 2005.

[19] Magnus, Jan R, and Heinz Neudecker. "Matrix differential calculus with applications in statistics and econometrics." (John Wiley \& Sons) 1995.

[20] Markowitz, Harry. "Portfolio selection*." *The journal of finance* (Wiley Online Library) 7, no. 1 (1952): 77-91.

[21] Maschek, Michael K. "Economic Modeling Using Evolutionary Algorithms: The Influence of Mutation on the Premature Convergence Effect." *Computational Economics* (Springer), 2015: 1-23.

[22] Meucci, Attilio. "Managing diversification." 2010.

[23] Oyenubi, Adeola. "Information theoretic measures of complexity and stock market analysis: Using the JSE as a case study" available at " http://open.uct.ac.za/handle/11427/10967 " (University of Cape Town) 2010.

[24] Rudin, Alexander M, and Jonathan S Morgan. "A portfolio diversification index." *The Journal of Portfolio Management* (Institutional Investor Journals) 32, no. 2 (2006): 81-89.

[25] Scrucca, Luca. "GA: a package for genetic algorithms in R." *Journal of Statistical Software* 53, no. 4 (2012): 1-37.

[26] Shannon, Claude E. "Bell System Tech. J. 27 (1948) 379; CE Shannon." *Bell System Tech. J* 27 (1948): 623.

[27] Statman, Meir. "How many stocks make a diversified portfolio?" *Journal of Financial and Quantitative Analysis* (Cambridge Univ Press) 22, no. 03 (1987): 353-363.

[28] Tang, Gordon YN. "How efficient is naive portfolio diversification? an educational note." *Omega* (Elsevier) 32, no. 2 (2004): 155-160.

[29] Van Emden, Maarten Herman. "An analysis of complexity." *MC Tracts* (Centrum Voor Wiskunde en Informatica) 35 (1971): 1-86.

[30] Van Heerden, JD, and Sonja Saunderson. "The Effect of the South African Market Concentration on Portfolio Performance." *Corporate Ownership \& Control*, 2008: 99.

[31] Willighagen, Egon, Michel Ballings, and Maintainer Michel Ballings. "Package 'genalg'." *Retrieved from*, 2015.

[32] Woerheide, Walt, and Don Persson. "An index of portfolio diversification." *Financial Services Review* (Elsevier) 2, no. 2 (1993): 73-85.

Figures 1, 2, 3 and 4: Simulation Result: Variation in Variance, ICOMP, PDI and HHI as number of stocks increases

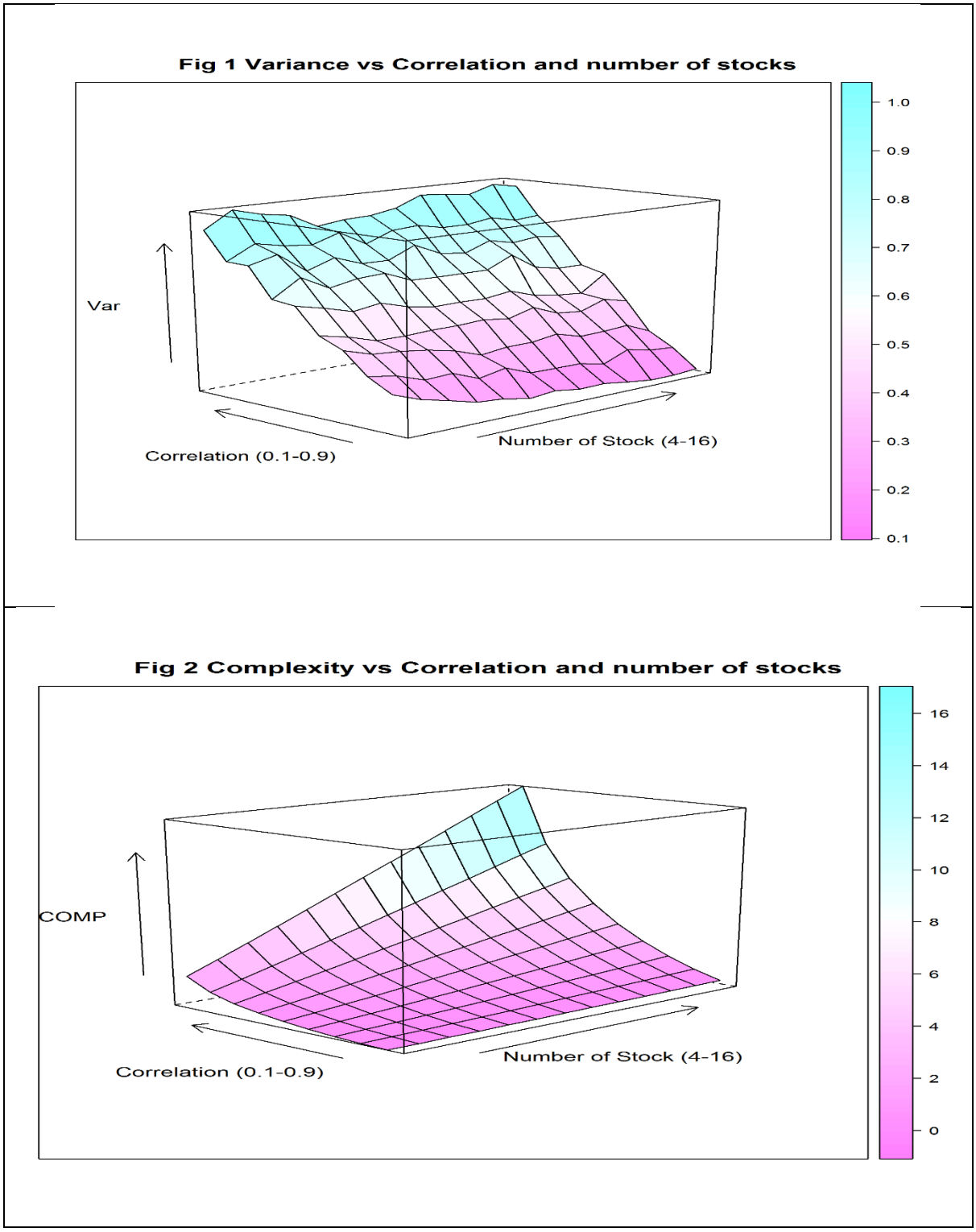


Fig 3 PDI vs Correlation and number of stocks

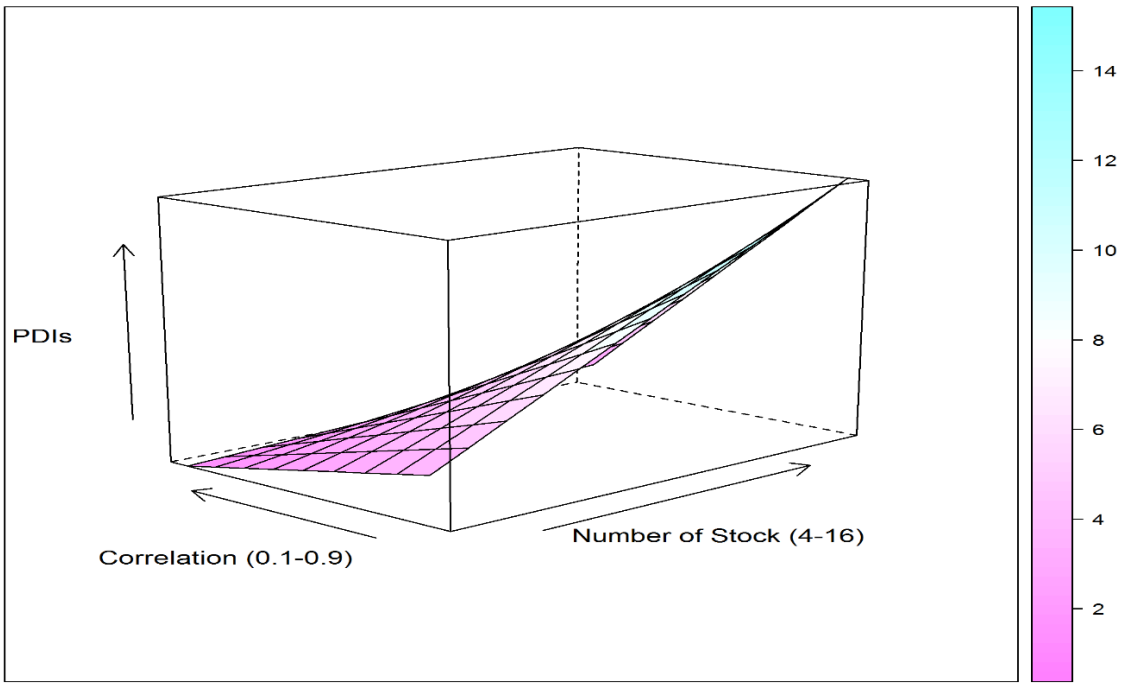
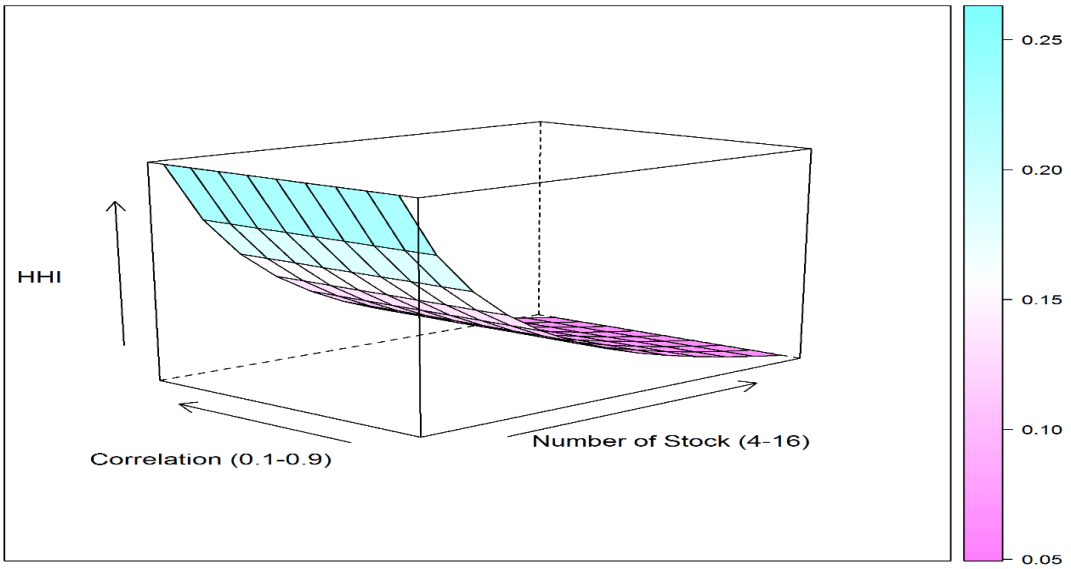
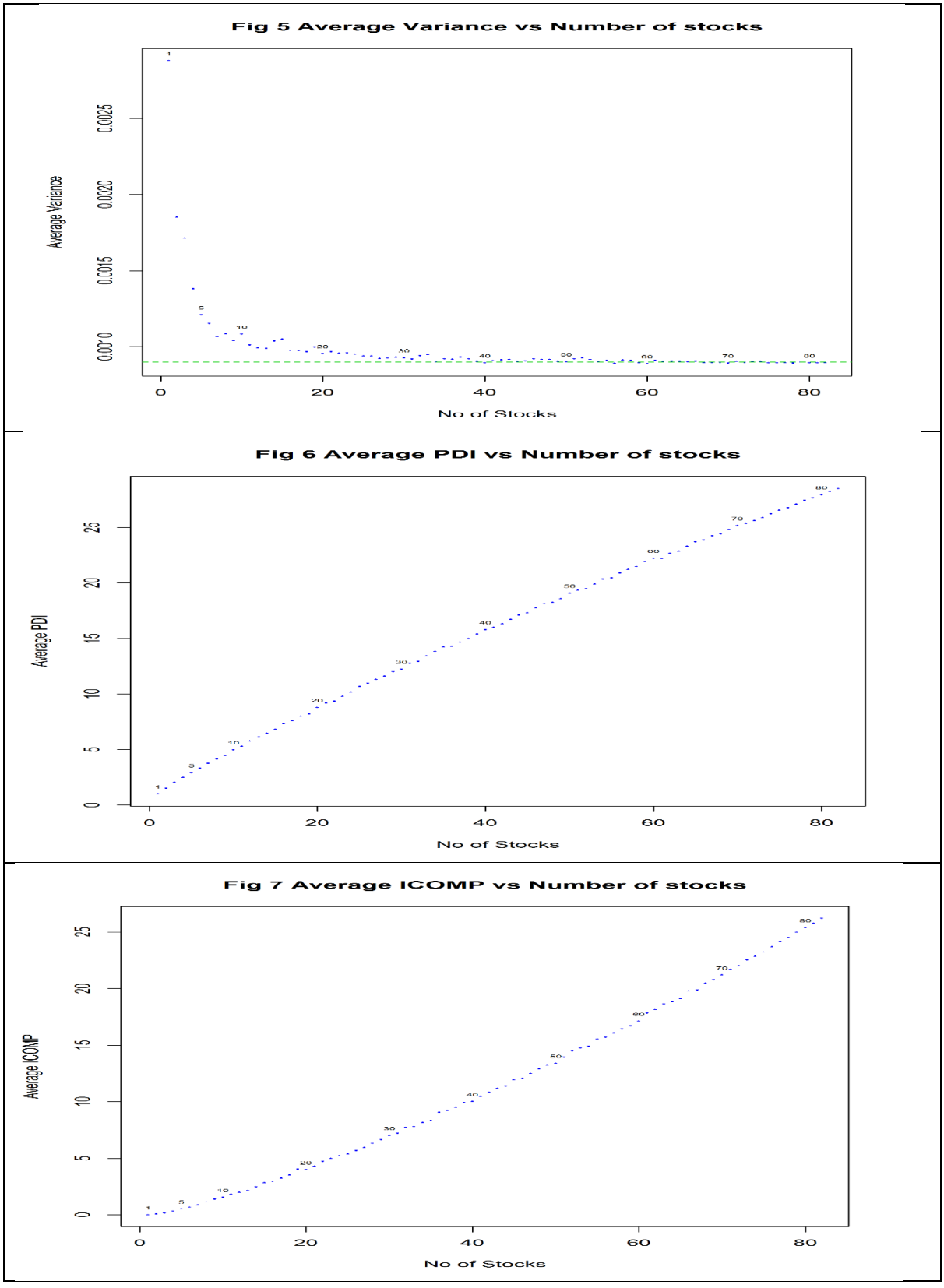


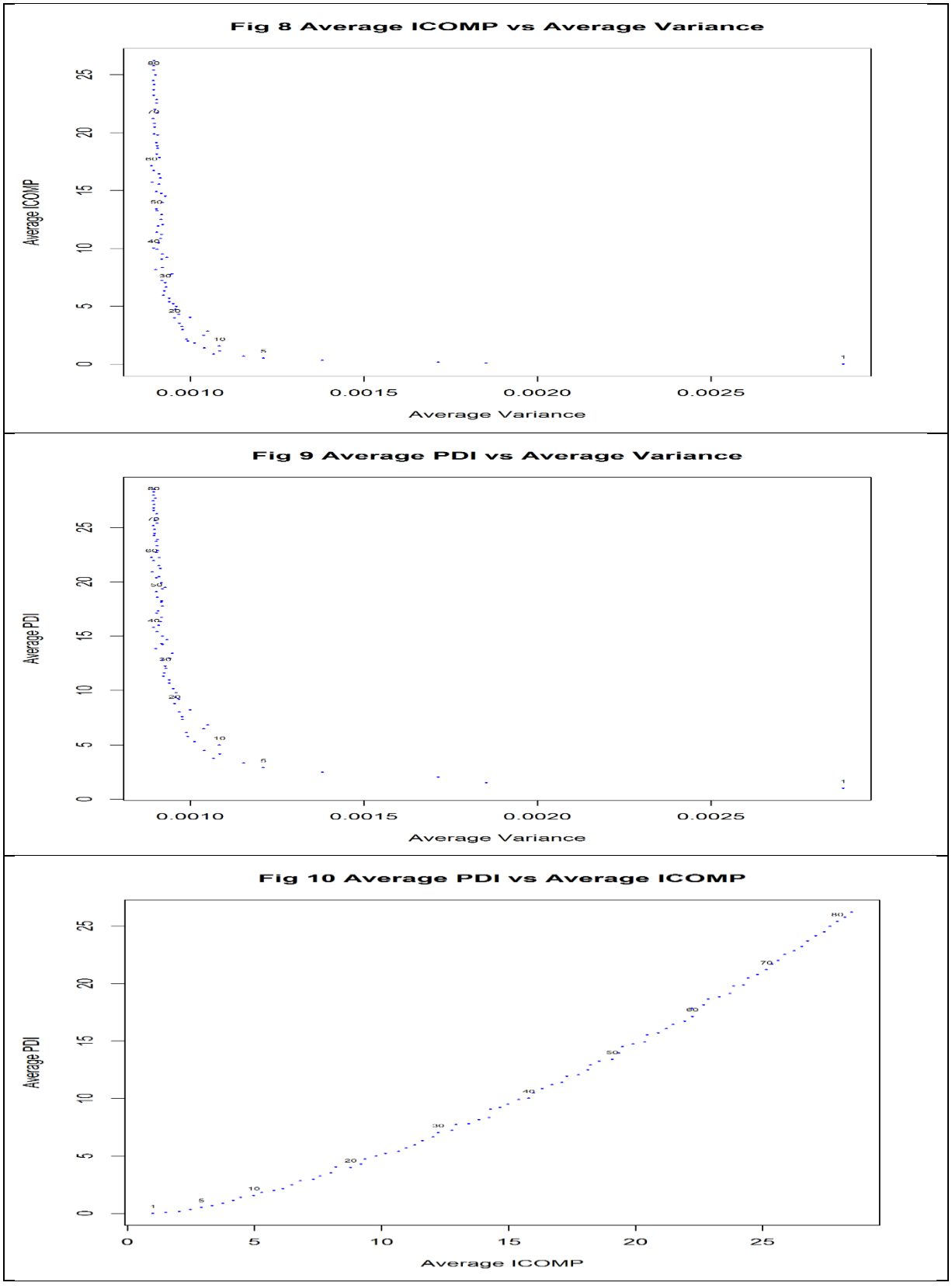
Fig 4 HHI vs Correlation and number of stocks



Figures 5,6 and 7: VARIANCE PDI ICOMP vs No of Stocks



Figures 8,9 and 10: Variation in PDI, ICOMP and Variance as number of stocks increase



Figures 11,12 and 13: Change in Average PDI vs Change in ICOMP, Average PDI vs Average ICOMP and Average variance vs Average ICOMP

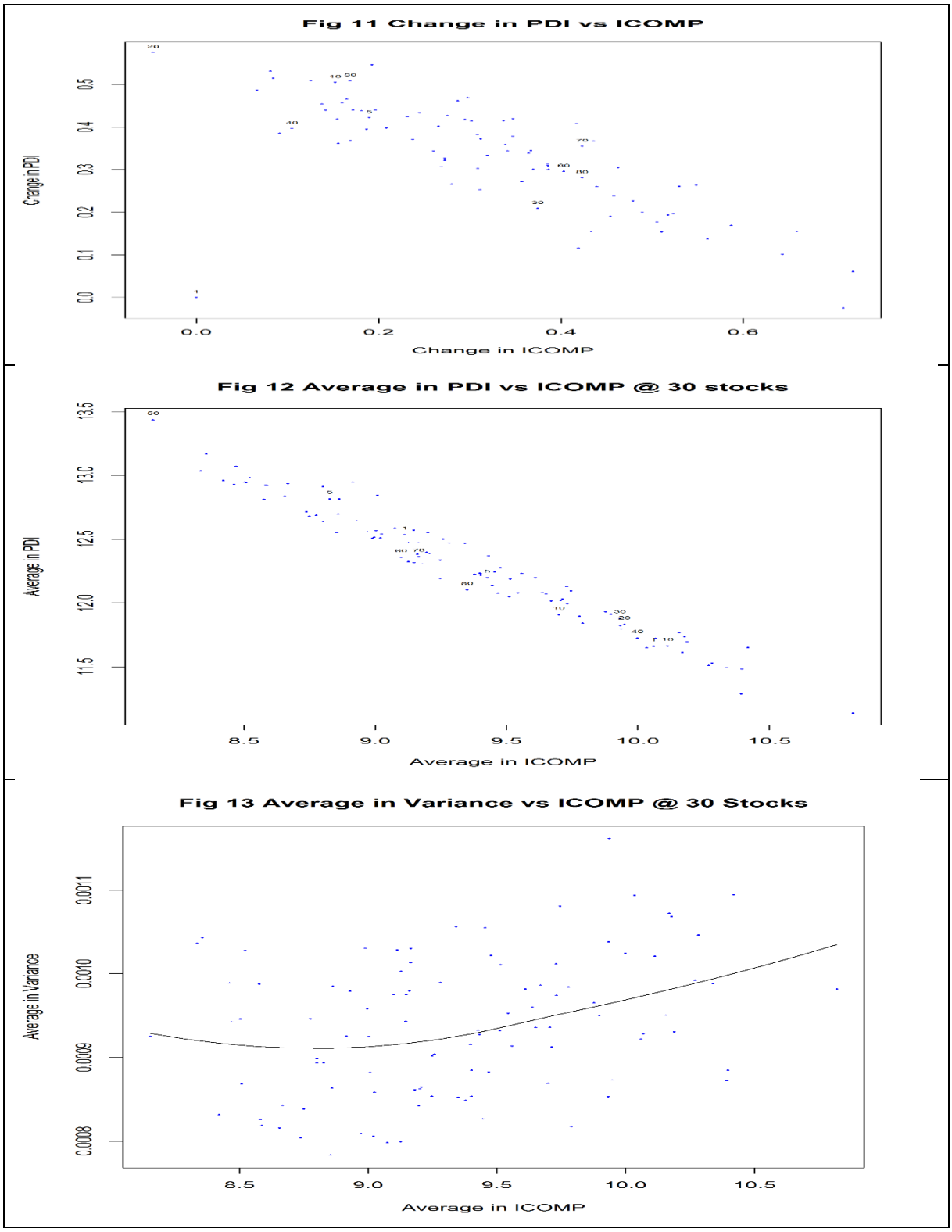


Figure 14: Difference between PDI and ICOMP

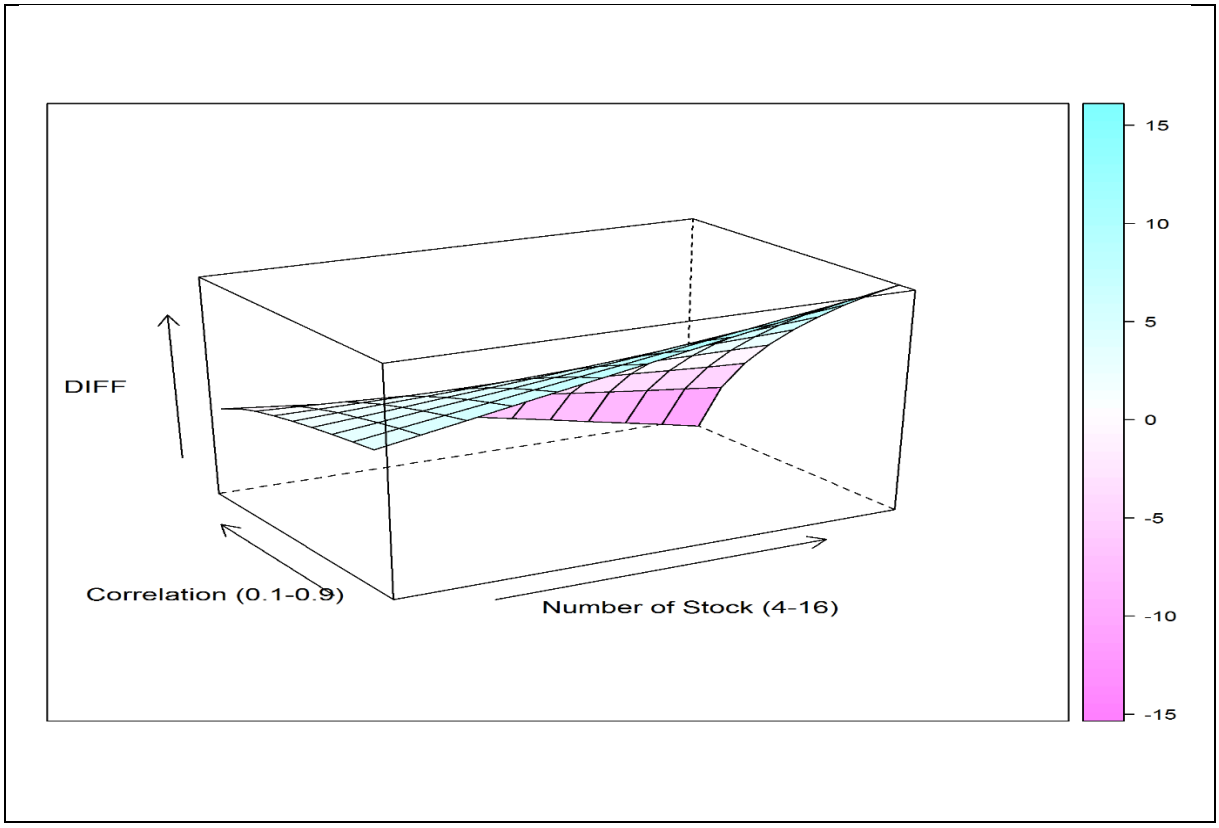


Figure 15: ICOMP vs MICOMP and MDI vs PDI

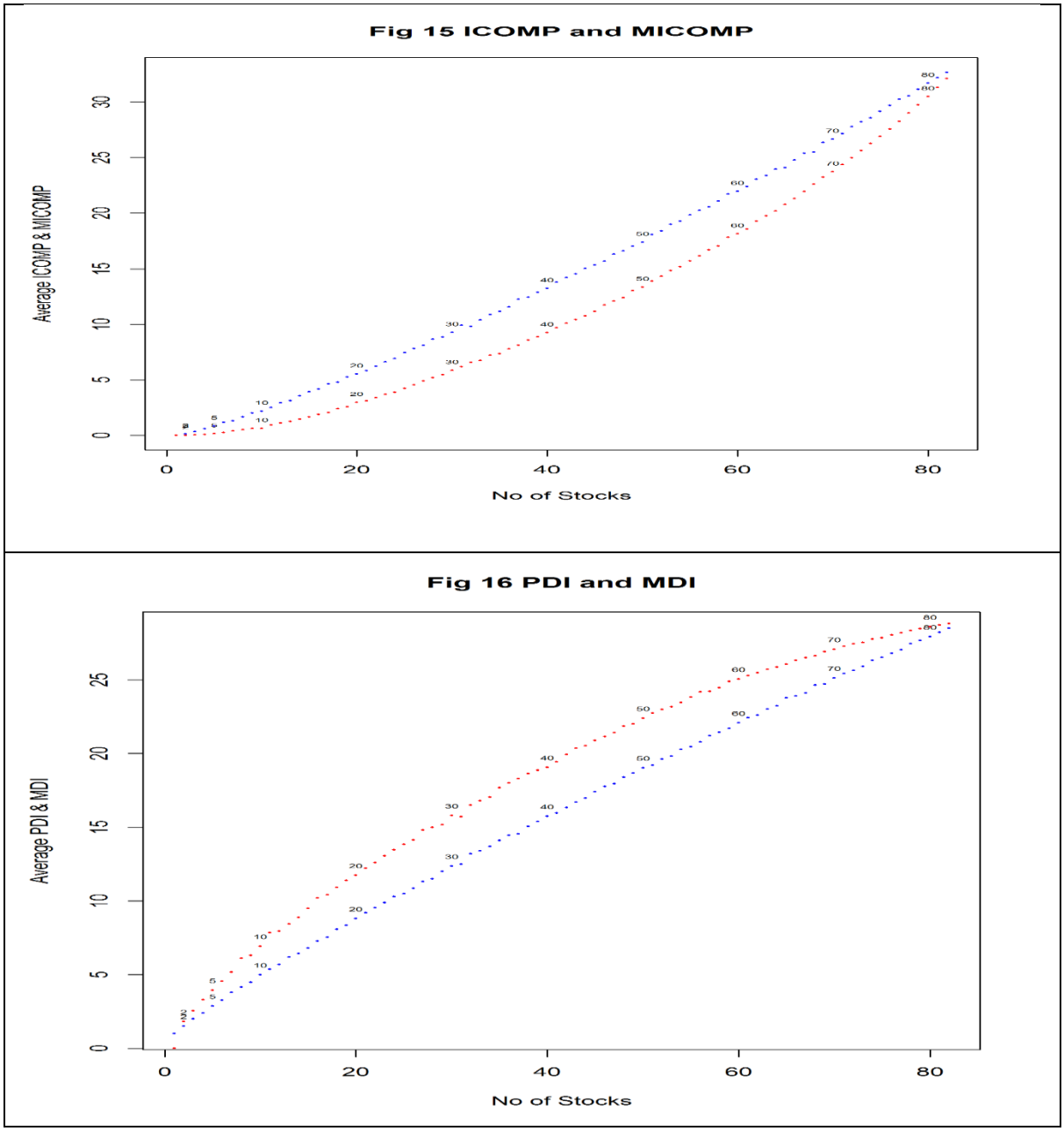


Figure 17: Trade-off between Diversification (PDI) and Complexity (ICOMP)

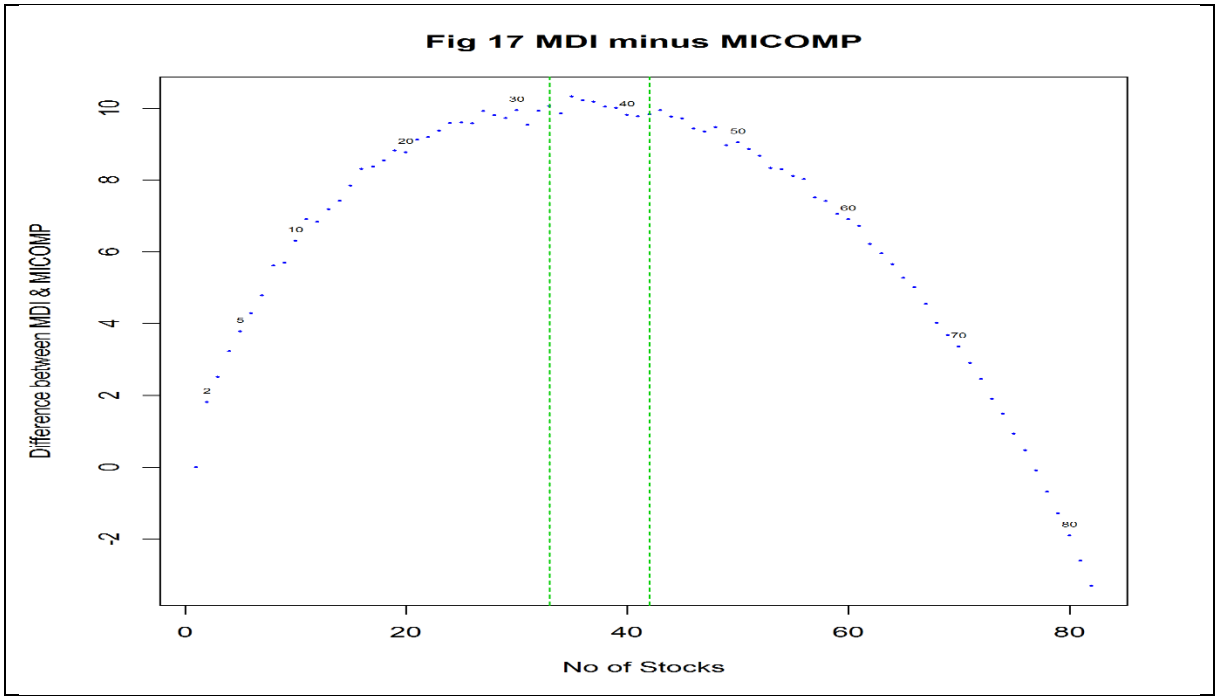
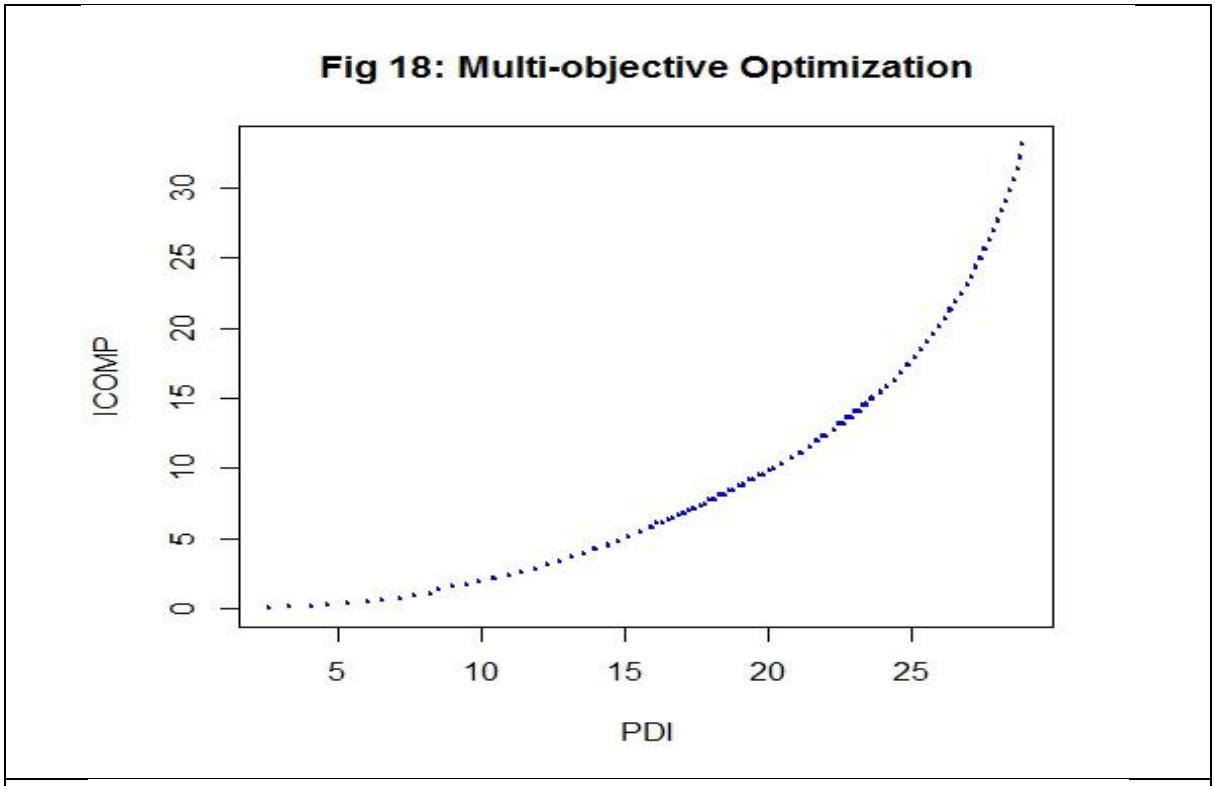
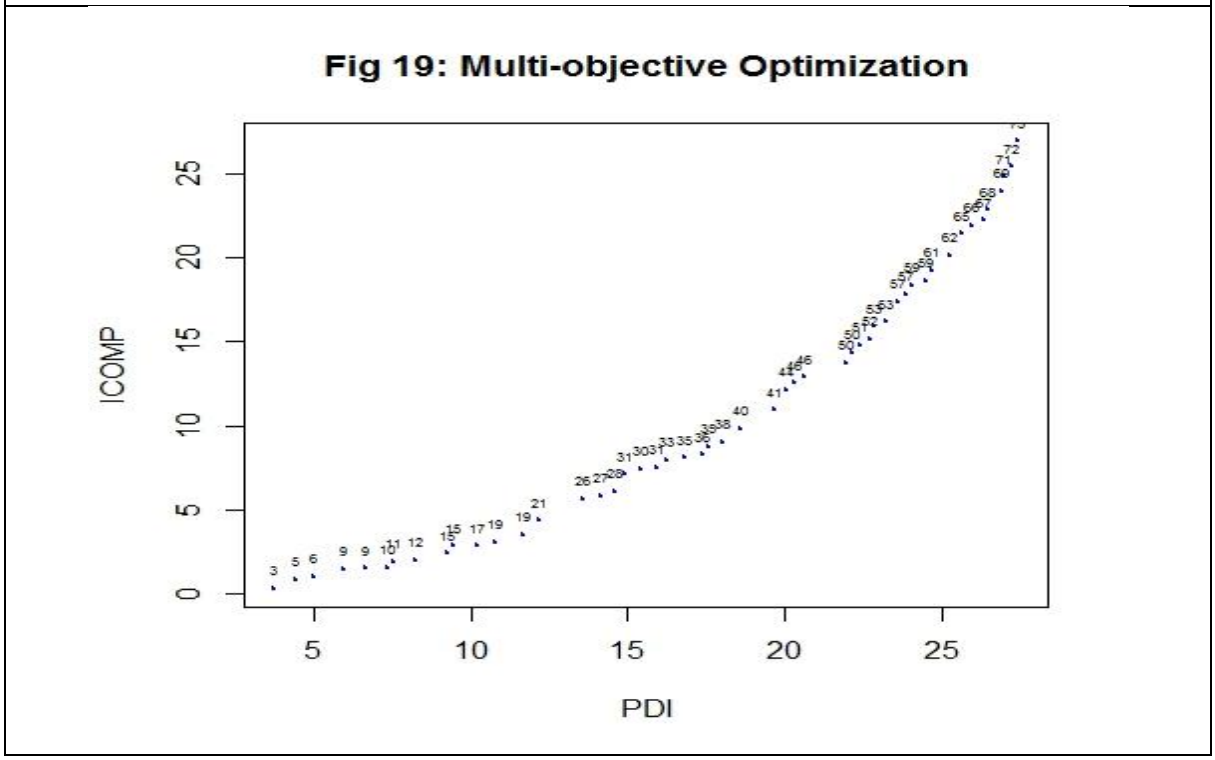


Figure 18 & 19: Multi-objective Pareto optimal portfolios





Appendix

	SYMBOL	COMPANY		SYMBOL	COMPANY
1	AAPL	Apple Inc.	**43**	KLAC	KLA-Tencor Corp.
2	ADBE	Adobe Systems	**44**	LBTYA	Liberty Global plc
3	ADI	Analog Devices	**45**	LLTC	Linear Technology Corp.
4	ADP	Automatic Data Processing Inc.	**46**	MAT	Mattel Inc.
5	ADSK	Autodesk Inc.	**47**	MCHP	Microchip Technology
6	AKAM	Akamai Technologies Inc	**48**	MDLZ	Mondelez International, Inc.
7	ALTR	Altera Corp.	**49**	MNST	Monster Beverage Corporation
8	ALXN	Alexion Pharmaceuticals, Inc	**50**	MSFT	Microsoft Corp.
9	AMAT	Applied Materials	**51**	MU	Micron Technology
10	AMGN	Amgen	**52**	MXIM	Maxim Integrated Products, Inc
11	AMZN	Amazon Corp.	**53**	MYL	Mylan Inc.
12	ATVI	Activision Blizzard, Inc.	**54**	NFLX	Netflix, Inc.
13	BRCM	Broadcom Corporation	**55**	NTAP	NetApp
14	CA	CA, Inc.	**56**	NUAN	Nuance Communications, Inc
15	CELG	Celgene Corp.	**57**	NVDA	Nvidia Corporation
16	CERN	Cerner Corporation	**58**	ORLY	O'Reilly Auto Parts
17	CHKP	Check Point Software Technologies Ltd	**59**	PAYX	Paychex Inc.
18	CHRW	C. H. Robinson Worldwide	**60**	PCAR	PACCAR Inc.
19	CMCSA	Comcast Corp.	**61**	PCLN	The Priceline Group Inc
20	COST	Costco Co.	**62**	QCOM	QUALCOMM Inc.
21	CSCO	Cisco Systems	**63**	REGN	Regeneron Pharmaceuticals, Inc.
22	CTSH	Cognizant Technology Solutions	**64**	ROST	Ross Stores Inc
23	CTXS	Citrix Systems	**65**	SBAC	SBA Communications Corp.
24	DLTR	Dollar Tree, Inc.	**66**	SBUX	Starbucks Corp.
25	DTV	DIRECTV Group Inc.	**67**	SHLD	Sears Holdings Corporation
26	EBAY	eBay Inc.	**68**	SIAL	Sigma-Aldrich
27	EQIX	Equinix, Inc	**69**	SIRI	Sirius XM Holdings Inc.
28	ESRX	Express Scripts	**70**	SNDK	SanDisk Corporation
29	EXPD	Expeditors Int'l	**71**	SPLS	Staples Inc.
30	FAST	Fastenal Co	**72**	SRCL	Stericycle Inc
31	FFIV	F5 Networks, Inc.	**73**	STX	Seagate Technology Public Limited Company
32	FISV	FIserv Inc.	**74**	SYMC	Symantec Corp.
33	FOSL	Fossil Group, Inc.	**75**	TXN	Texas Instruments
34	FOXA	Twenty-First Century Fox, Inc.	**76**	VOD	Vodafone Group Plc
35	GILD	Gilead Sciences	**77**	VRTX	Vertex Pharmaceuticals Incorporated
36	GMCR	Keurig Green Mountain, Inc.	**78**	WDC	Western Digital
37	GOLD	Randgold Resources Limited	**79**	WFM	Whole Foods Market, Inc
38	GOOG	Google Inc.	**80**	WYNN	Wynn Resorts Ltd

39	GRMN	Garmin Ltd.	**81**	XLNX	Xilinx Inc
40	HSIC	Henry Schein, Inc.	**82**	XRAY	Dentsply Intl
41	INTU	Intuit Inc.	**83**	YHOO	Yahoo Inc.
42	ISRG	Intuitive Surgical Inc.			
	Data downloaded using “get.hist.quote” command in R, spans Oct 10 2005 to Nov 25, 2013				