DEPARTMENT OF ECONOMICS AND FINANCE

COLLEGE OF BUSINESS AND ECONOMICS

UNIVERSITY OF CANTERBURY

CHRISTCHURCH, NEW ZEALAND

Stock Selection with Principal Component Analysis

Libin Yang[1], William Rea[2], Alethea Rea[3]

WORKING PAPER

No.3/2015

ISSN 1179-3228
Department of Economics and Finance
College of Business and Economics
University of Canterbury
Private Bag 4800, Christchurch
New Zealand

Stock Selection with Principal Component Analysis

Libin Yang[1], William Rea[2], Alethea Rea[3]

February 3, 2015

Abstract: We propose a stock selection method that is based on a variable selection method used with principal component analysis. We applied our method to stocks in the ASX200 and show that a portfolio of as little as 15 stocks can closely replicate the behaviour of the index. We show that the number of stocks required to form a diversified portfolio is not constant across time but varies with market conditions.

Keywords: Principal component analysis, stock selection, diversification, stock portfolios

JEL Classifications: G11

Acknowledgements:

1. Department of Economics and Finance, University of Canterbury, Christchurch, New Zealand
2. Department of Economics and Finance, University of Canterbury, Christchurch, New Zealand
3. Data Analysis Australia, Perth, Australia

*Corresponding Author: bill.rea@canterbury.ac.nz,

1 Introduction

PCA is one of the best-known descriptive techniques in multivariate analysis. Its range of applications has expanded with the advent of computers and it has been used in a wide variety of areas for the last 50 years (Jolliffe, 1986). The ability of PCA to decompose interrelated variables into uncorrelated components makes it attractive to use in analyzing the complex structure of financial markets. It has been applied to produce market indices (Feeney and Hester, 1967), identify common factors in bond returns (Driesson et al., 2003; Pérignon et al., 2007), to the study of market cross-correlation and systemic risk measurement (Kritzman et al., 2011; Zheng et al., 2012; Billioand et al., 2012) and identification of major risk components in a stock market (Kim and Jeong, 2005).

Most works have only discussed the theoretical framework of applying PCA in portfolio management, few have actually looked into its performance. Our research focuses on the practical application of PCA to portfolio management. Specificially we address the questions of

1. How many stocks does it take to create a diversified portfolio?

2. How can we identify which stocks to hold?

To the best of our knowledge, no similar work has been done based on the Australian market.

This first of these questions has a long history and a large literature. Lowenfeld (1909) discussed the benefits of diversification and sometimes is considered the first rigorous academic discussion of diversification though among market practitioners the benefits of diversification were known much earlier. Usually the scientific study of diversification is traced to Markowitz (1952). Markowitz argued that instead of looking at single security alone, one should be concerned with portfolios as a whole. The reason for including securities that have low correlations or even negative correlations is that they will eliminate some risk which has become known as idiosyncratic or diversifiable risk.

The question of how best to create a well-diversified portfolio is one which has many answers. The existence of many kinds of low-cost index funds provide a means for investors to hold a diversified portfolio. However, it is estimated that only about 11% of all funds invested in the stock market are held via index funds (The Economist, 2014), though that figure is growing.

It has long been argued that it is not necessary to include all constituents within a index to obtain the same level of diversification as the index itself. Indeed, Jacob (1974) pointed out that investors can reduce idiosyncratic risk significantly if they choose their securities judiciously. Conventional wisdom has it that the benefits of diversification are virtually exhausted when a portfolio contains a high enough number of stocks. Of course, that poses the question of how many stocks are enough.

Many researchers have based their studies on random selection and/or spreading the investments across industry groups while randomly selecting from within the groups (Statman, 1987; Domian et al., 2003, 2007). For randomly selected stocks, all stocks are assumed to be equally valuable. If randomly selected from within industry groups, it assumed that all stocks in the same industry are equivalent from an investment viewpoint. Even when one has found the number of stocks that exploit all the diversification benefits, it is nearly impossible to replicate the best combination of stocks that has the promised diversification because stocks do not have same mean return, variance and covariance. Blume and Friend (1978) reported that the actual diversification in 70 percent of the investors in their study was much lower than the number of securities in the portfolio suggested. It is very unlikely that investors were randomly selecting stocks, rather they had preferences for certain types of stocks. It is these preferences which made their portfolios under diversified.

Evans and Archer (1968) reported that approximately 10 randomly chosen stocks would be adequate to diversify a portfolio. They observed that the benefit of diversification decreased as the number of stocks held increased. Their conclusion has been cited in many textbooks widely used by finance students, see, for example, Gup (1983), Stevenson and Jennings (1984), Reilly (1985) and Francis (1986). Newbould and Poon (1993, 1996) followed Evans and Archer (1968)'s approach of comparing increasing portfolio size with its variance and claimed that just 8 to 20 stocks was enough to fully obtain the benefit of diversification. However, Statman (1987) compared the cost and benefit of diversification and reported that the number of randomly chosen stocks that make a well diversified portfolio was at least 30 using the data available in mid-1980s. When Statman (2004) used the same approach and more available data, he then concluded that the break even point, where the marginal benefit was equal to the marginal cost, exceeded 300 stocks.

Subsequently, many others have reported different numbers of stocks needed to diversify a portfolio using risk measurements other than variance. Domian et al.

(2003) reported that in order to avoid a significant shortfall risk, no less than 60 randomly chosen stocks were required. According to Domian et al. (2007), shortfall risk reduction continued as the number of randomly chosen stocks was increased, even above 100 stocks.

At the end of their analysis the above researchers were providng investors with a recommended minimum number of stocks in a portfolio which would serve as a measure of diversification. However, this was problematic. If, in an ideal world, all stocks had same mean, variance and covariance, the number of stocks in a portfolio would be the key variable for estimating the reduction in variance (Frahm and Wiechers, 2011). In reality, such assumptions do not hold. But even if investors randomly choose stocks to add to a portfolio, when the number of stocks required is reached, it may not have the promised diversification if the chosen stocks are more correlated than expected. This means that finding the number of stocks needed to diversify a portfolio may be useful from a theoretical point of view but remain impractical because one may not know which stocks should be held.

The use of PCA deals with the problem associated with randomly choosing stocks. Rudin and Morgan (2006) applied PCA to measure diversification quantitatively and tested equal-weighted portfolios of stocks in the S&P100 index and reported that a pool of 40 randomly selected stocks was approximately as diversified as only 20 truly independent components. PCA provides us with a way to identify uncorrelated risk sources in the market and pick stocks from those different risk sources, the resulting portfolio size is more meaningful from the point of view of diversification than a size reported on the basis of random or industry group selection.

Below we propose a stock selection method that picks stocks based on their correlation structure. The selected stocks are used to describe the original data set and represent the risk sources inherent in the data set.

An additional problem in trying to answer the question of how many stocks are enough is that the market connectedness does not stay constant over time. Markets become more tightly coupled in volatile periods and the level of diversification provided by a portfolio with same stocks would change over time. In particular it would become less diversified precisely when the protection provided by diversification is most needed, namely in times of market turmoil (Fenn et al., 2011; Kritzman et al., 2011; Billioand et al., 2012; Zheng et al., 2012).

Campbell et al. (2001) recognised this point and reported that the number of stocks needed to achieve a certain level of diversification was not the same in the 1963-85 period and the 1986-97 period.

The remainder of this paper is organized as follows; Section (2) describes the data, gives some descriptive statistics and the methods used, Section (3) outlines our stock selection method, Section (4) presents the results and discussion, Section

(5) our conclusions.

2 Data, Descriptive Statistics and Method

2.1 Data

Our research is based on the Australian market. The main index for the market is the ASX200, which is a market capitalization weighted index of the 200 largest shares by capitalization listed on the Australian Securities Exchange. The index in its current form was created on 31 March 2000. We investigated the constituents of the ASX200 index from inception to February 2014. Figure (1) shows the index values over the full study period. The ASX200 index is a capitalization index and so does not adjust for dividends. In our research we calculated the returns for all constituents which included the dividends paid.

There was a high frequency of stocks that were added to or deleted from the index from time to time, so we identified all stocks which had been in the ASX200 at any time during the whole study period. After adjusting for mergers, acquisitions, and name changes we obtained a final data set of 524 unique stocks. We obtained daily closing prices and dividends for each stock from the SIRCA database[1]. All the prices and dividends were adjusted to be based on the AUD. The return was calculated in the following steps:

1. We created a new variable associated with each stock called the Dividend Factor. We started with a factor of 1 and every time a dividend was paid we multiplied the Dividend Factor,

$$\text{Daily Dividend Factor}_i(t) = \left\{ \begin{array}{ll} 1 & \text{if no dividend} \\ 1 + \frac{D_i(t)}{P_i(t)} & \text{if dividend} \end{array} \right\}$$

$$\text{Cumulative Dividend Factor}_i(t) = \prod_{j=1}^{t} (\text{Daily Dividend Factor}_i(t))$$

   where $D_i(t)$ is the dividend for stock i at time t, $P_i(t)$ is price for stock i at time t in units of one trading day.

2. We adjusted the price series with the dividend factor, the adjusted price was calculated by

$$\text{PNEW}_i(t) = P_i(t) \times \text{Cumulative Dividend Factor}_i(t).$$

[1] `http://www.sirca.org.au/`

3. The return series for a given stock i was calculated as

$$\mathrm{R}_i(t) = \frac{\mathrm{PNEW}_i(t+1) - \mathrm{PNEW}_i(t)}{\mathrm{PNEW}_i(t)}.$$

2.2 Descriptive Statistics

In Figure (2) we investigated the characteristics of the ASX200 return data and the time series plot of the ASX200 percentage return, a box plot as well as the 100 largest absolute returns and a Quantile-Quantile plot compared to the normal distribution. These graphs were generated with functions in the package `fBasics` (Wuertz et al., 2013) within `R` (R Core Team, 2014).

We found evidence of volatility clustering: "large changes tend to be followed by large changes, of either sign, and small changes tend to be followed by small changes" (Mandelbrot, 1963). From the two plots in the left-hand panel, we observe that most large absolute returns occurred during the 2008 financial crisis. The ASX200 index level continued to change significantly until the end of 2009. There was also a cluster of large returns at the end of 2011, this is when the Australian stock market was affected as investors responded to America's credit downgrade, the European sovereign debt crisis, and fears over the global economy. Moreover, in the box plot and QQ plot, the ASX200 daily returns are skewed to the left and the heavy tails are evident. We further calculated the skewness and kurtosis of the ASX200 daily returns. There were -0.383 and 5.657 respectively, clearly indicating a heavy tail.

2.3 Methods

PCA can be applied to either a correlation matrix or a covariance matrix but there are some problems associated with using a covariance matrix. If there are large differences between the variances of variables, then using a covariance matrix will result in the low numbered principal components being dominated by variables that have a large variance. This will impede getting useful information from a PCA in some cases (Jolliffe, 1986). We found this was the case for our data so all PCAs reported here were done on correlation matrices of the return series as calculated above.

For most parts of our research, a rolling window approach was applied. We extracted a set of stocks that had been in the index at any time and for which there was complete return information for the whole study period, and there were 156 such stocks. The remaining 368 stocks were either listed after April 2000 or delisted before February 2014.

We used the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy (Kaiser, 1970; Kaiser and Rice, 1974) to test the shortest length of sliding window that a

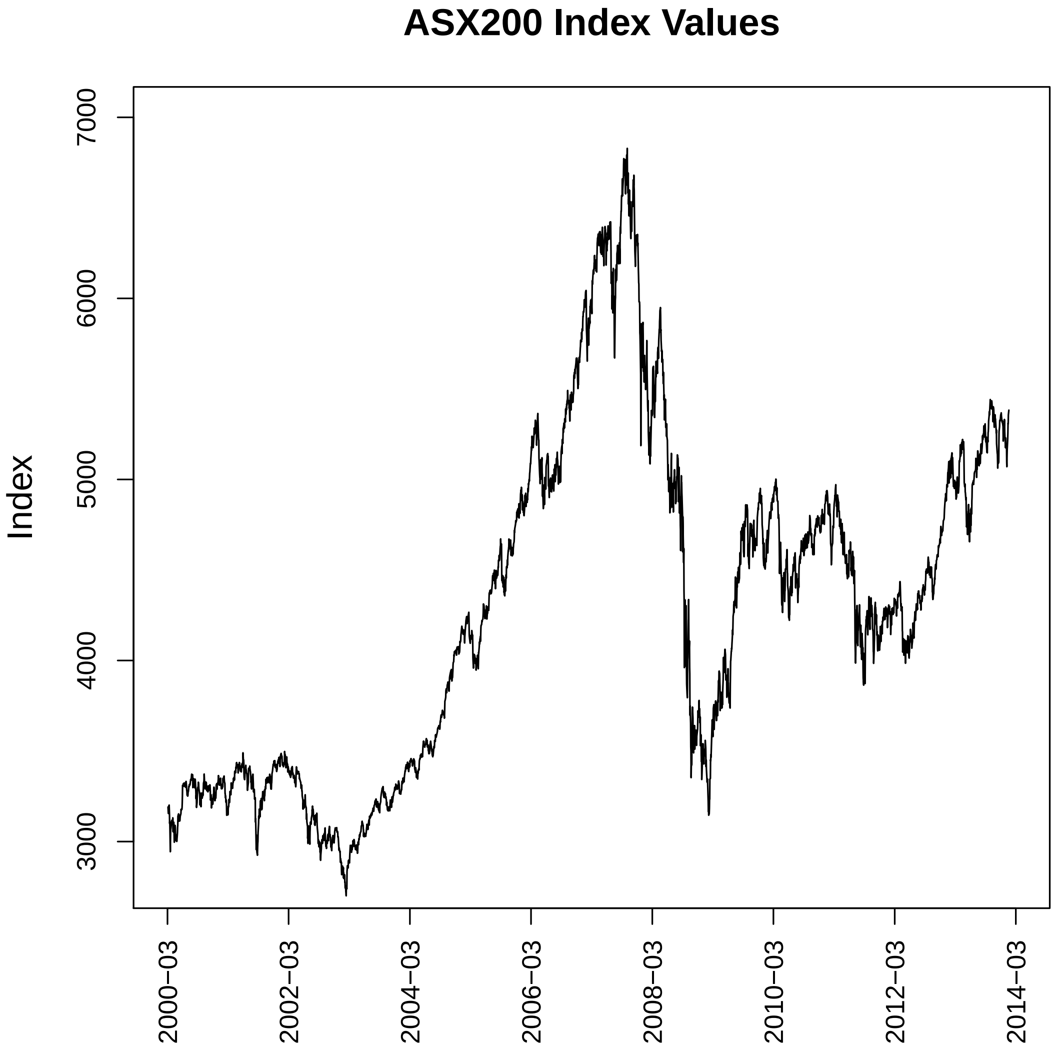
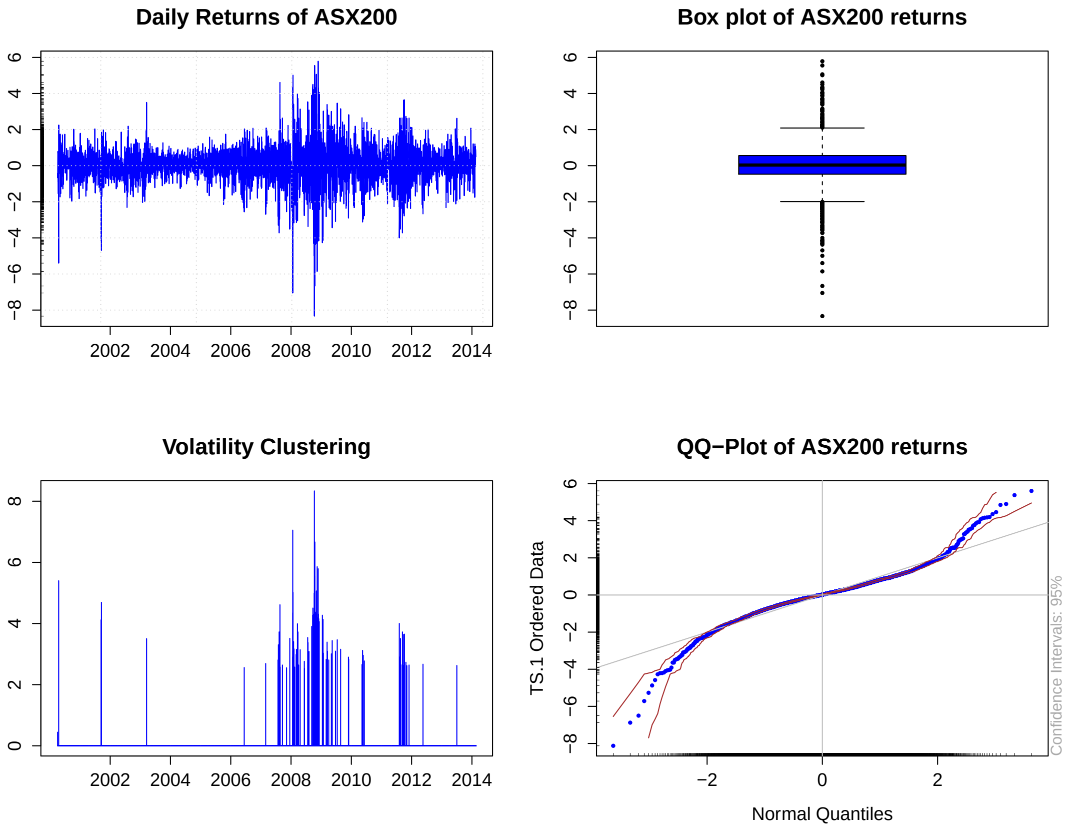


Figure 1: The ASX200 index values for the study period.

Figure 2: Stylized facts for ASX200.



Daily Returns of ASX200

Box plot of ASX200 returns

Volatility Clustering

QQ−Plot of ASX200 returns

PCA could be efficiently applied to. The KMO statistic compares the value of correlations between stocks to those of the partial correlations. If the investigated stocks share more common variation, the KMO will be close to 1. On the other hand, a KMO near 0 indicates the PCA will not extract much useful information. A KMO value of 0.5 is the smallest KMO value that is considered acceptable to do a PCA. The KMO test was performed using functions in the `R` package `psych` (Revelle, 2014).

We calculated the KMO statistic in rolling windows of different sizes for the 156 stocks which had complete data. We settled on a window size of two years or 504 trading days. Within these windows the KMO values ranged between 0.62 and 0.95. These values indicate that a PCA can be usefully applied to the data.

Correlation matrices were generated with the `cor` function in base `R`. PCAs were carried out using the function `eigen` in base `R`. The efficient frontier plots were made with functions in `fPortfolio` (Wuertz et al., 2014) and random portfolios generated with function in `rportfolios` (Novomestky, 2012).

3 Stock Selection Method

Jolliffe (1986) pointed out that if a data set can be successfully described by a smaller number of principal components, then it will always be true that it can be replaced by a subset of the original variables. When the number of variables in a data set is large, it is often the case that many variables contain repeated information. So it will be the case that a subset of variables contains a large proportion of the information available in the full data set. We show below with properly choosing stocks, a much smaller portfolio will closely resemble the ASX200 index in terms of the fluctuation in portfolio value.

We followed Jolliffe (1986) and used the variable selection method that he claimed to retain the "best" subsets more often than other methods considered. This method is related to Kaiser's rule (Kaiser, 1960) which retains principal components from a correlation matrix with an eigenvalue greater than one.

The selection procedure is as follows:

1. Apply PCA to the correlation matrix of a stock market.

2. Associate one stock with the highest coefficient in absolute value with each of the last m_1 principal components that have eigenvalues less than a certain level, l, which we call the deletion criteria, then delete those m_1 stocks. For example, one can use Kaiser's rule because in the case of a correlation matrix, a principal component with an eigenvalue less than 1 contains less information than one of the original variables.

3. A second PCA is performed on remaining stocks. The same procedure is applied that associates one stock with each m_2 principal components that have an eigenvalue less than l, and delete those m_2 stocks.

4. The procedure is repeated until no further deletions are considered necessary based on a stopping criteria. One can decide to stop the selection procedure based on the eigenvalue of the last principal component. For example, the stopping criteria can be; delete stocks until the principal components of the retained stocks have eigenvalues not less than 0.7.

To better understand the selection procedure we describe it in detail when applied to the 156 stocks with complete data for the whole period with a deletion criteria of 1 and a stopping criteria of 0.7.

First deletion cycle We performed a PCA on the correlation matrix of the 156 stocks and there were 107 principal components with eigenvalues lower than 1. We found the stocks with the highest coefficient in each of the 107 principal components and there were 84 unique stocks. Note that some stocks have the highest coefficient in more than one principal component. We removed these 84 stocks from the sample.

Second deletion cycle We performed a PCA on the 72 retained stocks. The eigenvalue of the last principal component was 0.49, which was lower than the stop criteria of 0.7. There were 47 principal components which had eigenvalues lower than 1 with 40 unique stocks associated with the components. We deleted these 40 stocks from the sample.

Third deletion cycle We performed a PCA on the 32 retained stocks. The last principal component had an eigenvalue of 0.64 which was lower than the stop criteria of 0.7. We again deleted the stocks associated with the principal components which had eigenvalues lower than 1 and 15 stocks were retained.

Fourth cycle A PCA was performed on the 15 retained stocks and the last eigenvalue was 0.77, higher than the 0.7 stop criteria. We stopped the deletion and there were 15 stocks selected after three cycles of deletion.

The idea behind this method is that low eigenvalue principal components are often associated with near-constant relationships among a subset of variables (Jolliffe, 1986, p43). If such variables are detected and deleted, little information will be lost. With each step of the deletion procedure, the eigenvalues of the new set of principal components will converge. In the example discussed above, most of principal components from the selected 15 stocks have eigenvalues close to each other. The second largest eigenvalue was 1.12 and the smallest is 0.77. This means

No. Stocks Retained	Maximum Correlation
15	0.17
32	0.27
72	0.45
156	0.71

Table 1: The maximum correlation between stocks in each of the four deletion cycles described in Section (3).

each principal component contains a similar amount of information as one individual stock. The principal components obtained from the selected 15 stocks were approximately the same as the original 15 stocks. This is the case of when there is low correlation among the original stocks a PCA extracts little useful information.

One can control the deletion speed by adjusting the deletion criteria. Jolliffe (1986) suggested that deleting principal components that have eigenvalue less than 1 is too aggressive and likely to result in a loss of useful information, a more conservative level is 0.7. Thus we could have set the deletion criteria to 0.7 which would have slowed the deletion process. The combination of deletion criteria and stopping rule determines how many stocks are retained.

4 Results and Discussion

4.1 Stocks with Complete Data

We have described the stock selection procedure with a deletion criteria of 1 and a stopping criteria of 0.7 on 156 stocks for the whole study period in detail in Section (3). We further investigated the performance of the selected stocks. We will discuss the three sets of stocks which were selected from the three levels of the deletion cycles. Recall that a deletion criteria of 1 and stopping criteria of 0.7 required three deletion cycles and retained 72, 32, and 15 stocks at the end of one, two and three cycles respectively.

As can be seen in Table (1) when more stocks were retained, the maximum correlation in the portfolio increased. With each step of deletion procedure, stocks with the highest correlations with the other stocks were deleted.

Figure (3) presents the efficient frontier constructed from our 15 selected stocks. The red dot is the mean and standard deviation of an equally weighted portfolio of the 15 selected stocks. The blue dots are the means and standard deviations of 1000 equally weighted randomly selected portfolios of 15 stocks.

For the random portfolios, the stocks were selected from the 156 stocks in our data set without replacement. It is clear that except for four portfolios, all random portfolios of 15 stocks lie inside the achievable region, which is inside the efficient frontier in Figure (3). This means there will be at least one portfolio constructible from the selected 15 stocks that has the mean and volatility corresponding to each of the 996 random portfolios. That is, with the 15 stocks selected by PCA we were able to replicate 99.6 percent of portfolios with 15 randomly selected stocks. Thus the 15 stocks selected from our method explain the original 156 stocks well.

We also constructed an efficient frontier based on one of the random 15 stock portfolios for comparison purposes and this is presented in Figure (4). As can be seen, a lot of the random portfolios lie outside the efficient frontier in Figure (4).

Unsurprizingly, we find that the selected 32 stocks describe the original data set better than the selected 15 stocks. In Figure (5), all the random portfolios lie in the achievable region. Compared to the 32 stocks selected by our method, the 32 stocks randomly picked from the full data set can not achieve all the means and volatilities corresponding to the 1000 random portfolios (see Figure 6). The selected 72 stocks from our method is not superior to the 72 stocks randomly selected in terms of describing the full data set (see Figures 7 and 8). Moreover, by comparing the mean and standard deviation of portfolios of selected stocks to the random portfolios of the same size, we find that the 15-stock portfolio and, to a lesser extent, the 72-stock portfolio lie toward the middle of the random portfolio cluster. In contrast, the selected 32-stock portfolio lies in the left edge of the random portfolio cluster. Intuitively, the portfolio of selected 32 stocks tends to have lower risk for the given level of return or higher return for the given level of risk compared to the random portfolios.

We have found that each of the three different numbers of selected stocks all explain well the original 156 stocks. When comparing the risk and return of portfolios of selected stocks to the random portfolios, 32-stock portfolio was slightly better than either the 15 or 72 in the sense that in Figures (3) to (8) the 32 stock portfolio was closer to the left edge of the cloud of 1000 simulated portfolios than the 15 stock portfolio but achieves this at a lower transaction cost than the 72 stock portfolio. Curiously, the 32 stock portfolio had a lower weekly volatility than the 72.

We further compared the risk and return of the three selected portfolios to try to find the point where the benefits of diversification are virtually exhausted. The portfolio of 32 stocks has slightly reduced the risk and had higher returns compared to the portfolio of 15 stocks. When the portfolio size was increased to 72, the return increased but the risk was higher compared to the portfolio of 32 stocks. All three portfolios lie close to the global minimum variance point, which is the lowest possible variance a portfolio of 156 stocks. We conclude that 15 stocks are not enough to diversify a portfolio and the 32 stocks selected by our method is where all the diversification benefits are exploited when using the whole

Figure 3: The efficient frontier constructed from the 15 stocks selected by PCA, the mean and standard deviation of equally weighted portfolio of the selected 15 stocks and the means and standard deviations of 1000 equally weighted random portfolio of 15 stocks selected from the 156 stocks in our data set. The inset is the cloud of 1000 random portfolios and the 15 stock portfolio selected by PCA. All the returns are on a weekly basis.

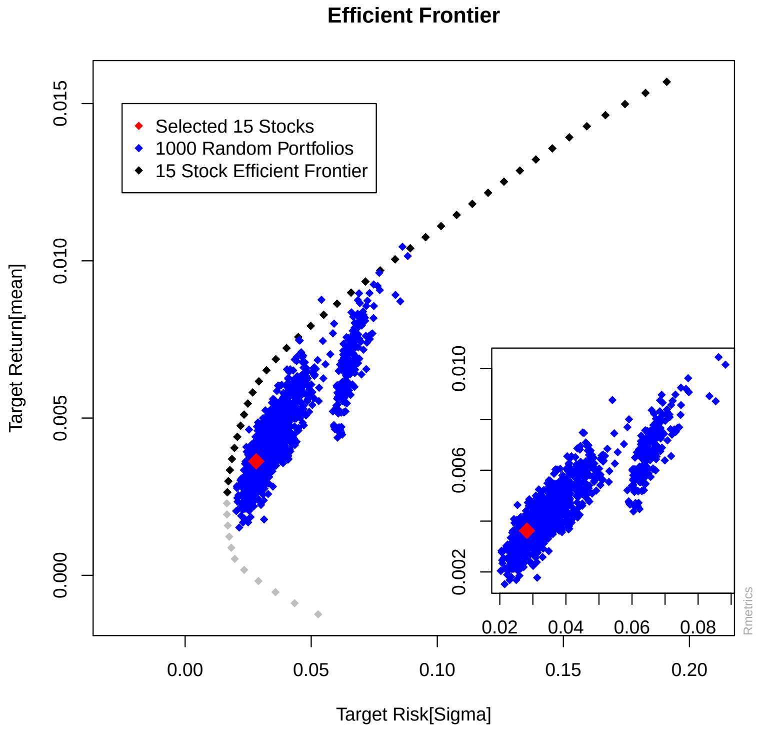


Figure 4: The efficient frontier constructed from 15 randomly selected stocks, the mean and standard deviation of equally weighted portfolio of the 15 stocks selected by PCA and the means and standard deviations of 1000 equally weighted random portfolio of 15 stocks selected from the 156 stocks in our data set. The inset is the cloud of 1000 random portfolios and the 15 stock portfolio selected by PCA. All the returns are on a weekly basis.

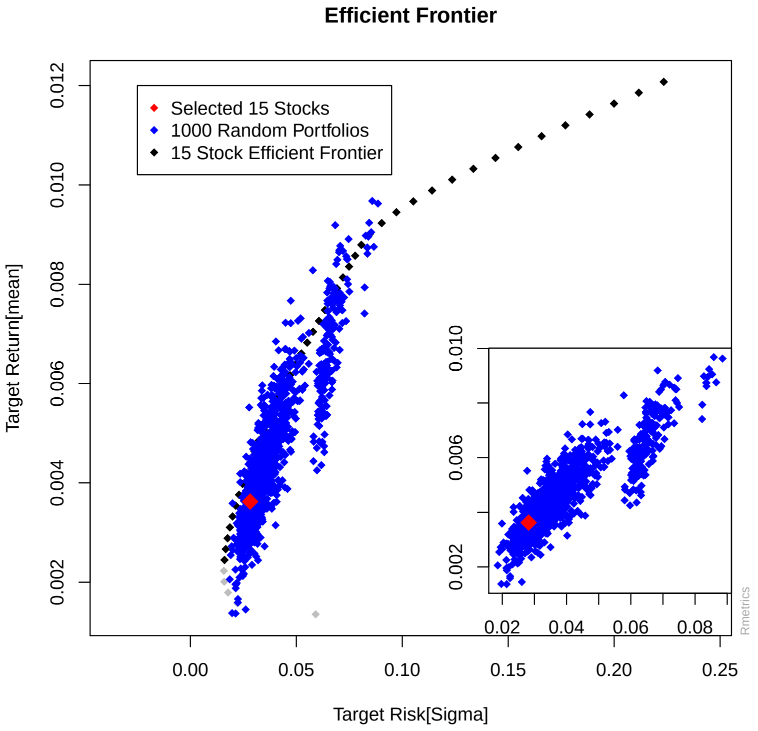


Figure 5: The the efficient frontier constructed from the 32 stocks selected by PCA, the mean and standard deviation of an equally weighted portfolio of selected 32 stocks and the means and standard deviations of 1000 equally weighted random portfolios of 32 stocks selected from the 156 stocks in our data set. The inset shows the cloud of 1000 random portfolios and the 32 stock portfolio selected by PCA. All the returns are on a weekly basis.

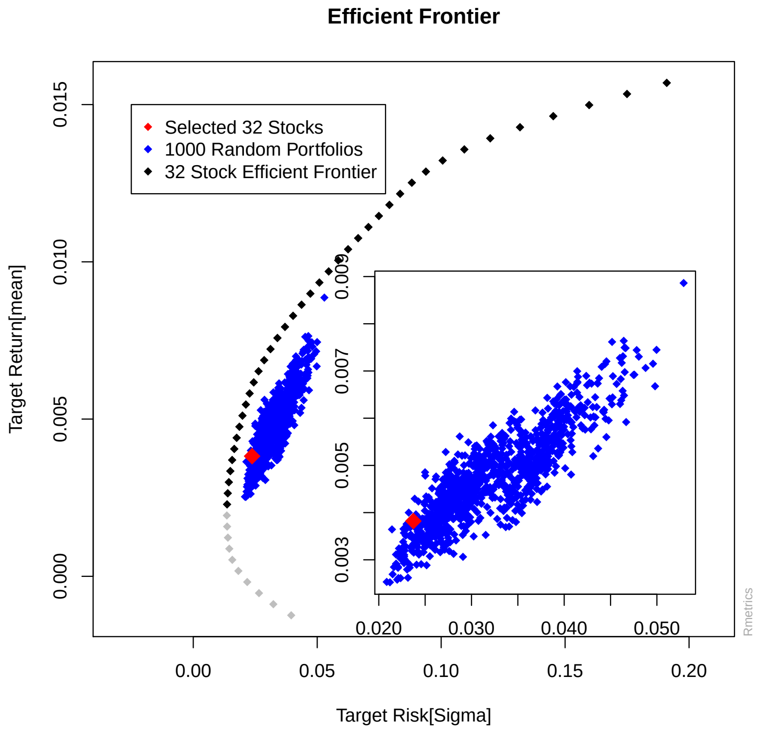


Figure 6: The the efficient frontier constructed from 32 randomly selected stocks, the mean and standard deviation of an equally weighted portfolio of 32 stocks selected by PCA and the means and standard deviations of 1000 equally weighted portfolios of 32 randomly selected stocks from the 156 stocks in our data set. The inset shows the cloud of 1000 random portfolios and the 32 stock portfolio selected by PCA. All the returns are on a weekly basis.

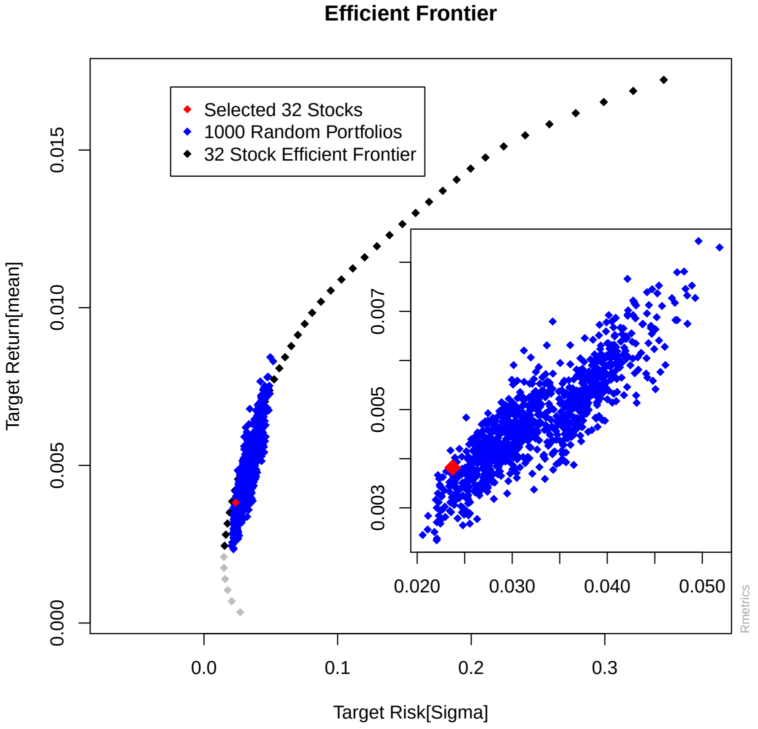


Figure 7: The efficient frontier constructed from the 72 stocks selected by PCA, the mean and standard deviation of an equally weighted portfolio of 72 stocks, and the means and standard deviations of 1000 equally weighted portfolios of 72 randomly selected stocks from the 156 stocks in our data set. The inset shows the cloud of 1000 random portfolios and the 72 stock portfolio selected by PCA. All the returns are on a weekly basis.

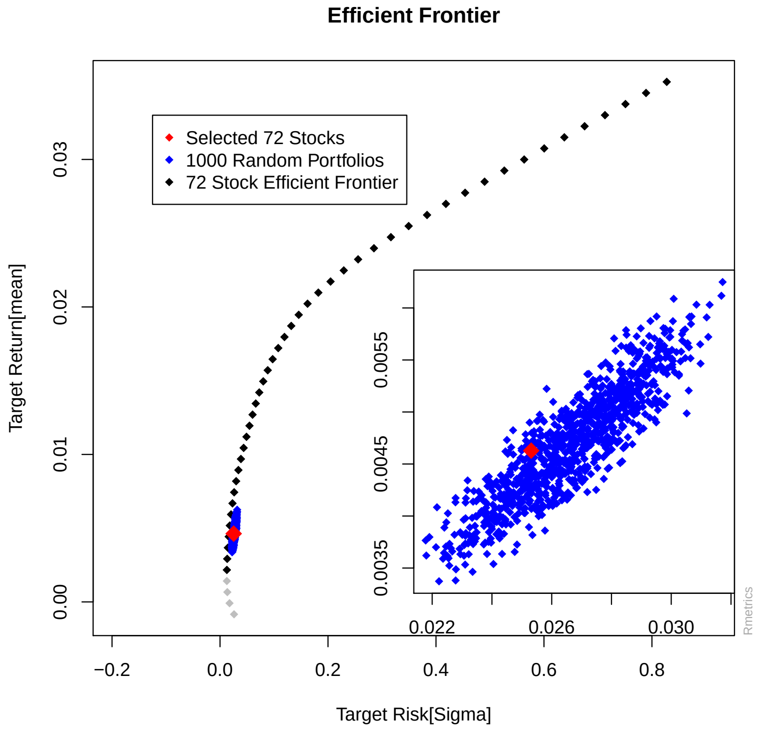
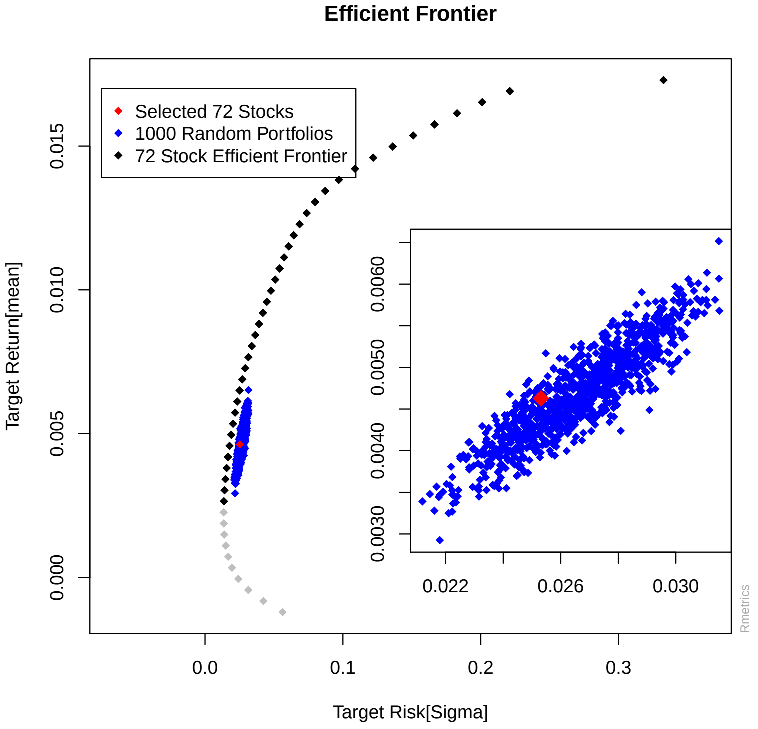


Figure 8: The efficient frontier constructed based on 72 randomly selected stocks by PCA, the mean and standard deviation of equally weighted portfolio of the 72 stocks selected by PCA and the means and standard deviations of 1000 equally weighted random portfolio of 72 stocks selected from the 156 stocks in our data set. The inset shows the cloud of 1000 random portfolios and the 72 stock portfolio selected by PCA. All the returns are on a weekly basis.



study period for the investigation. Further spreading the portfolio's investment to include 72 stocks is superfluous diversification and should be avoided.

Table (2) and Table (6) lists the 15 and 32 selected stocks together with their industry information respectively. The stocks selected were spread across almost all industries. There are a total of 10 industries represented in the ASX200 index based on the Industry Classification Benchmark (ICB). The 15-stock portfolio included nine out of the 10 industries while the 32 stock portfolio contained all industries. We found that when the number of stocks was increased from 15 to 32, the stocks added were also spread over all industries. Moreover, we noticed that in both the 15 and the 32 selected stocks, major companies such as BHP and RIO in Basic Materials, the four big banks (ANZ, CBA, NAB, WBC) in the Financials, and WPL and STO in the Oil & Gas industry, were not selected.

There are two explanations for their omission; one statistical, one financial. These stocks were highly correlated and so had high weightings in the last few principal components. This ensured their elimination early in the deletion cycle. A financial explanation is that the major companies were exposed to multiple risk sources and so they have a tendency to move with the the broad market thus providing little benefit for diversification. The stock selection procedure tends to select stocks that represent the uncorrelated risk sources in the market. These major companies are correlated with multiple other companies because they are diversified within their respective sectors.

4.2 Stock selection using full data set

The correlations between stocks changed over time and this affects the number of stocks selected. Based on this, we suspected that during the periods of a more connected market, there should be less risk sources. This means one should expect a smaller number of selected stocks are required to describe the market.

Our final test of stock selection was to examine the performance of the selected stocks compared to the ASX200 index value. We found that, in general, the fluctuation of the ASX200 index value can be replicated with a much smaller portfolio, but not a portfolio of constant size.

All the tests in Section (4.1) were based on the 156 stock data set. We were also concerned that stock selection was sensitive to the selection pool. With different stocks available to be chosen, the selection procedure may result in a very different set of stocks. In order to better compare with ASX200 index, using more complete constituents was considered more appropriate. So we divided the whole study period into seven subperiods, each with a length of two years[2] (around 504 trading

[2] The years were from a March 31 start date.

Table 2: The 15 stocks that were selected from the 156 stocks used for the whole study period, based on a deletion criteria of an eigenvalue 1 and stop criteria of 0.7.

Stock Code	**Industry**
MAH	Basic Materials
TRY	Basic Materials
AVG	Consumer Goods
ELD	Consumer Goods
MTS	Consumer Services
VRL	Consumer Services
DJW	Financials
IBC	Financials
IOF	Financials
RHC	Health Care
AJL	Industrials
HIL	Industrials
AUT	Oil & Gas
SMX	Technology
HTA	Telecommunications

days)[3], except for the last sub period which is less than two years and only had 472 trading days. We extracted the stocks that had complete returns information in the relevant periods. Table (4) in Appendix A summarizes the number of stocks in selection pool in each two year sub period.

The results in Table (3) illustrates that the number of stocks needed to diversify a portfolio is not constant through time. With the number of major stock market risk sources changing, a portfolio can be considered diversified consistently only if it is adaptive to the change. In other words, the number of stocks included to diversify major risk sources should change based on the number of risk sources in the market. Thus, a portfolio that holds the same number of stocks or same constituents can only be the best combination to create a diversified portfolio at a single point of time. Holding more stocks than necessary when the number of major risk sources decreases is redundant. On the other hand, holding fewer stocks than required when the number of major risk sources increases means that the portfolio is under-diversified.

We performed in-sample and out-of-sample tests of stocks selected for each two year sub-period. We compared the portfolio value of stocks selected using the first years' data to the portfolio value of stocks selected using the second year. In financial terms this is the comparison of the portfolio which would have been held with the portfolio which should have been held. We also performed the KMO test on each two year sub-period data, and the shortest length of data to efficiently apply PCA was one year. Table (5) in Appendix A presents the KMO statistic of each year. From 2006, the KMO statistics were all above 0.7. There was only one year, 2004 to 2005, the KMO statistic went below 0.5, the lowest acceptable value. The portfolio construction was carried out in following steps:

1. Within each two year sub-period, the stock selection procedure described in Section (3) was applied to the first year and second year separately. This created two sets of selected stocks based on the first and second year's data respectively.

2. For each set of selected stocks, we constructed a portfolio that had equal investment in those stocks. We called the portfolios of stocks selected the "first period model portfolio" and the "second period model portfolio".

3. For both portfolios, and the ASX200, we assumed a $1 million investment based on the prices at the first day of second year. The first portfolio value will converge to $1 million at the first day of second year and diverge subsequently.

[3] All the study periods are two years exactly but the actual number of trading days may vary.

4. For second period model portfolio, we only computed the second year portfolio value.

The test results are presented in Figures (9) and (10). All graphs have the same vertical scale to aid comparisons between periods. The gray vertical line indicates the first day of the second year, where the portfolio value equal to $1 million. The right hand side of the gray line shows the portfolio value of stocks held, stocks that should have been held and the ASX200 index value. It is clear from the Figures that portfolios of selected stocks, regardless of the period from which they were selected, moved closely relative to the index. Especially in periods of 2004 to 2006 and 2012 to 2014, the trajectory of first period model portfolio and second period model portfolio approximately matched the index. Moreover, the selected stock portfolios consistently outperformed the index in most of the second periods. However, this is almost certainly because the model portfolios included dividends.

We found that the first period model in general more closely resembled the ASX200 index in out-of-sample testing than the second period model except for the period of 2008 to 2010. In those two years the second period model portfolio closely evolved with the index, the first period model portfolio was far more volatile in 2009. One explanation for this is that during the 2008 financial crisis, the market conditions changed significantly and the affects caused by the crisis lasted a long time. The first period in 2008 to 2010 is completely different from its second period. Consequently, the portfolio of stocks selected from the first period market conditions was not adapted well to the second period market conditions.

While the trajectory of selected stocks portfolios shows that the ASX200 index can be described by smaller portfolios, we next investigated the number of stocks that were selected in these portfolios. Table (3) presents the number of stocks selected in the first period and second period of each two years sub-period. For the first three two year sub-periods, the number of stocks selected was all above 20 and the maximum is 25. The difference between the first period and second period was not larger than two stocks. This minor difference also reflected in the selected portfolio values in Figures (9) and (10). The red line (second period model portfolio) and the blue line (first period model portfolio) are almost matched. The difference in the number of stocks selected between the first period and second period increased for the subsequent study periods. The trajectories of period two portfolios in the strong rise pre-2008 and in the post-crisis period are less similar. The number of stocks selected declined to below 20. This is the reflection of a more connected market. For the first year of the last study period, the number of stocks rose to 21 but in the second year this number decreased. Our results indicate that to adequately diversify a portfolio, one does not have to include all 200 stocks. A portfolio with about 20 stocks well described the 200 stock index.

Table 3: The number of stocks selected for each year. A deletion criteria of an eigenvalue 0.7 and stop criteria of 0.5 was used. The final column specifies how many stocks were common to both periods.

Study Period	1st period	2nd period	Common Stocks
2000-2002	21	20	2
2002-2004	23	25	2
2004-2006	22	21	0
2006-2008	18	14	2
2008-2010	13	17	2
2010-2012	19	12	1
2012-2014	21	17	3

For investors who want to buy individual stocks and replicate the fluctuation of the index, our method of stock selection provides a way to make this possible.

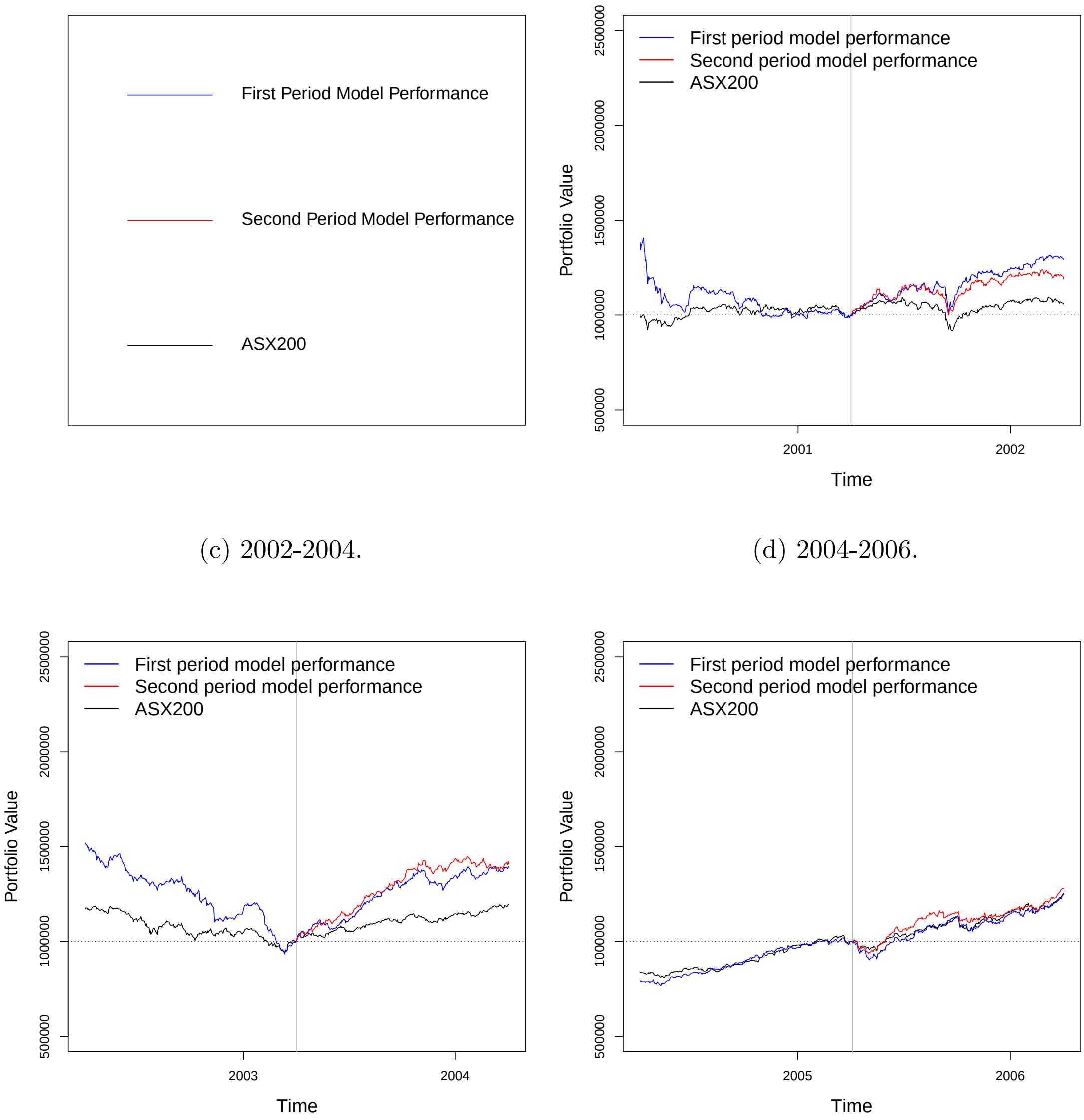
5 Conclusions

It is unwise to use a single selection rule for picking stocks to hold in a portfolio. The evidence presented here shows that the stocks selected through our method give a good level of diversification for the number of stocks held. In addition, the portfolios formed were able to replicate the index behavior well except during the period around the 2008 financial crisis.

There are several ways in which our selection method could be used. It is possible to generate a selection of stocks of any desired sze through careful choice of deletion and stopping rules. The evidence from Table (3) and Figures (3) through (10) suggests that the performance of the portfolio is not unduly sensitive to the stocks selected. Thus one way to use the selection method would be to generate a pool of potential investments which would be subject to further analysis before the final selection is made. For example, if an investor wished to hold a portfolio of, say, 15 stocks, the selection procedure could be used to generate, say, a pool of 30 potential stocks. The investor would apply their usual stock selection analysis to the 30 and on the basis of that analysis pick 15 for their portfolio. The way the pool is constructed almost certainly ensures that the final portfolio will have better diversification than if the whole market were considered.

Another possible use would be to compare an existing portfolio with the portfolio of the same size selected by our procedure. Again, the way our portfolio is constructed would almost certainly ensure that it was better diversified than that actually held by an investor except in the case that they were the same. The

Figure 9: In sample and out of sample test of portfolios of selected stocks against the ASX200 index value. Stock selection was based on deletion criteria of an eigenvalue of 0.7 and stop criteria of 0.5. for the years 2000 to 2006.



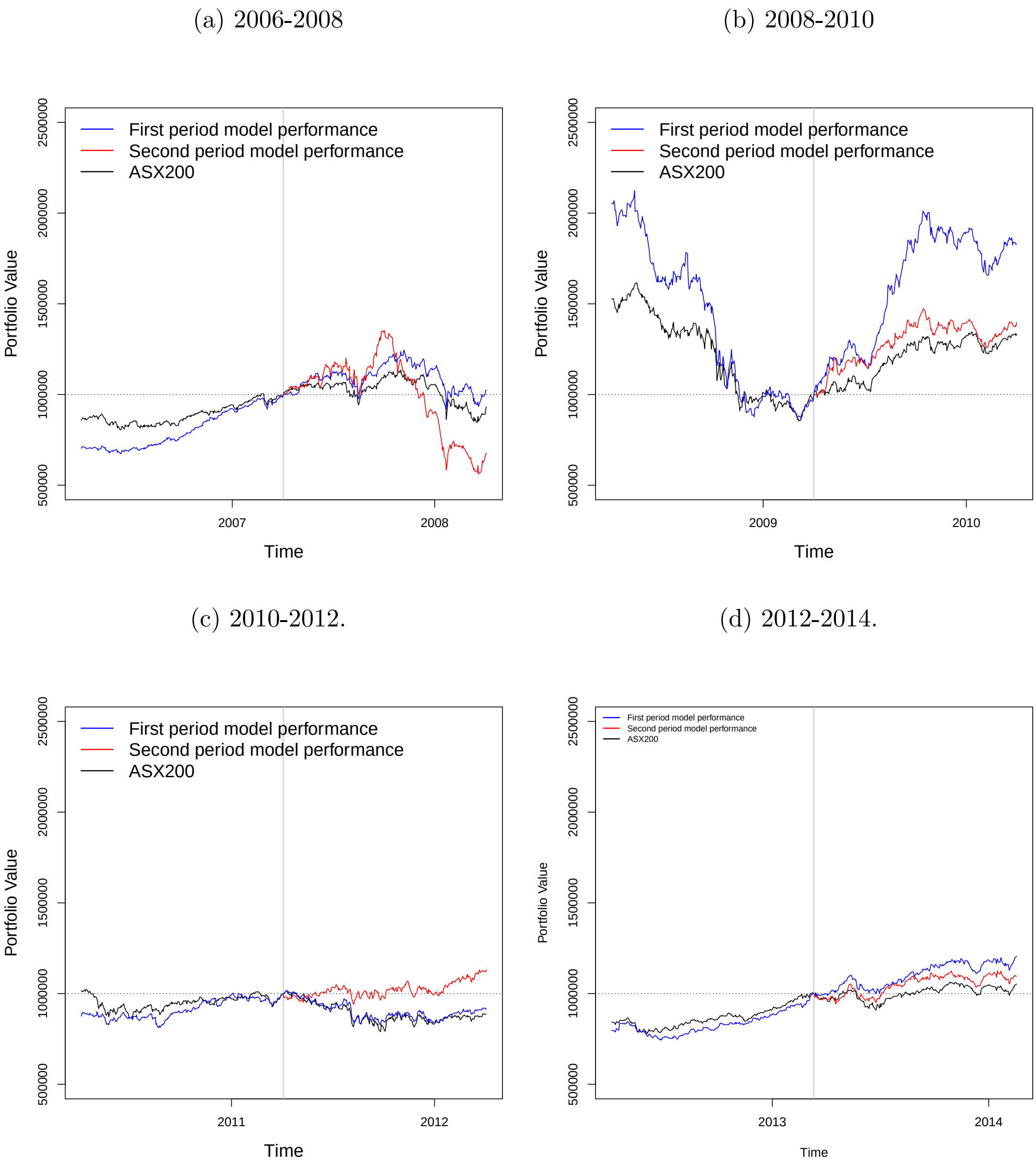
(a) Legend.

(b) 2000-2002.

(c) 2002-2004.

(d) 2004-2006.

Figure 10: In sample and out of sample test of portfolios of selected stocks against the ASX200 index value. Stock selection was based on deletion criteria of an eigenvalue of 0.7 and stop criteria of 0.5 for the years 2006 to 2014.


(a) 2006-2008

(b) 2008-2010

(c) 2010-2012.

(d) 2012-2014.

stocks which were in one portfolio but not the other could be subject to further analysis and a decision about whether to trade could be made. For example, in our results above the major mining, banking, and oil and gas stocks were all eliminated. Yet there would be few investors who would not hold one or more of these stocks. Because diversification is not the only consideration an investor has, the investor may choose to continue holding these stocks and accept the additional risk of doing so. But highlighting them for further analysis gives an investor the chance to reconsider, and perhaps reconfirm, their inclusion in the portfolio.

On the question of how many stocks are required to form a diversified portfolio we have found that there is no single number which can be offered as an answer. The answer depends both on the market conditions and on the method of selection. The method of selection we have proposed is likely to be the minimum number of stocks required to achieve a given level of diversification. Other selection methods will likely require more stocks. For example, when using our subset of 156 stocks, we found that 32 stocks were enough to form a well-divesified portfolio. If picking stocks randomly or from industry groups, then more than 32 stocks are likely to be required. Our results confirm previous studies which have shown that spreading stocks across industry groups is necessary.

To try to answer this question in greater detail requires much more extensive testing of different sized portfolios which result from different deletion and stopping criteria than we have done here and more detailed comparision with other selection methods.

A counter-intuitive result in Table (3) shows that as the correlations between stocks in the market rise, as they do in times of both euphoria and crisis, the number of stocks required decreases. This is a consequence of there being less diversification opportunities available in the market. To obtain better diversification an investor needs to add other asset classes.

While we have applied our selection method to portfolios of stocks, it is, in fact, quite general and will work with any set of investment opportunities provided the correlations between them can be estimated.

References

Billioand, M., M. Getmansky, A. W. Lo, and L. Pelizzon (2012). Econometric measures of connectedness and systemic risk in the finance and insurance sectors. *Journal of Financial Economics 104*, 535–559.

Blume, M. E. and I. Friend (1978). *The Changing Role of the Individual Investor: A Twentieth Century Fund Report* . New York: John Wiley & Sons.

Campbell, J., L. Martin, M. Burton, and Y. X. Xu (2001). Have Individual stocks become more volatile? An empirical exploration of idiosyncratic risk. *Journal of Finance 56*(1), 1–43.

Domian, D. L., D. L. Louton, and M. D. Racine (2003). Portfolio diversification for long holding periods: How many stocks do investors need? *Studies in Economics and Finance 21*, 40–64.

Domian, D. L., D. L. Louton, and M. D. Racine (2007). Diversification in portfolios of individual stocks: 100 stocks are not enough. *The Financial Review 42*, 557–570.

Driesson, J., B. Melenberg, and T. Nijman (2003). Common factors in international bond returns. *Journal of International Money and Finance 22*, 629–656.

Evans, J. L. and S. H. Archer (1968). Diversification and the reduction of dispersion: An empirical analysis. *Journal of Finance 23*, 761–767.

Feeney, G. J. and D. D. Hester (1967). *Risk Aversion and Portfolio Choice.* New York: Wiley.

Fenn, D. J., M. A. Porter, S. Williams, M. MacDonald, N. F. Johnson, and N. S. Jones (2011). Temporal evolution of financial-market correlations. *Physical Review E 84*, 026109.

Frahm, G. and C. Wiechers (2011). On the diversification of portfolios of risky assets. *Discussion paper in statistic and econometrics*. [online] `https://ideas.repec.org/p/zbw/ucdpse/211.html`.

Francis, J. C. (1986). *Investment: Analysis and Management* (Fourth ed.). New York: McGraw-Hill.

Gup, B. E. (1983). *The Basics of Investing* (Second ed.). New York: John Wiley & Sons.

Jacob, N. L. (1974). A Limited-Diversification Portfolio Selection Model for The Small Investor. *Journal of Finance 29*, 837–857.

Jolliffe, I. T. (1986). *Principal Component Analysis.* New York: Springer.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement 20*, 141–151.

Kaiser, H. F. (1970). A second generation little jiffy. *Psychometrika 35*(4), 401–415.

Kaiser, H. F. and J. Rice (1974). Little jiffy. *Educational and Psychological Measurement 34*(1), 111–117.

Kim, D.-H. and H. Jeong (2005). Systematic analysis of group identification in stocks markets. *Physical Review E 72*, 046133.

Kritzman, M., Y. Li, S. Page, and R. Rigobon (2011). Principal Components as a measure of systemic risk. *Journal of Portfolio Management 37*, 112–126.

Lowenfeld, H. (1909). *Investment, an Exact Science.* Financial Review of Reviews.

Mandelbrot, B. B. (1963). The variation of certain speculative prices. *Journal of Business 36*, 392417.

Markowitz, H. (1952). Portfolio Selection. *Journal of Finance 7*, 77–91.

Newbould, G. D. and P. S. Poon (1993). The minimum number of stocks needed for diversification. *Financial Practice and Education 3*, 85–87.

Newbould, G. D. and P. S. Poon (1996). Portfolio risk, portfolio performance, and the individual investor. *Journal of Investing 5*, 72–78.

Novomestky, F. (2012). *rportfolios: Random portfolio generation.* R package version 1.0.

Pérignon, C., D. R. Smith, and C. Villa (2007). Why common factors in international bond returns are not so common. *Journal of International Money and Finance 26*, 284–304.

R Core Team (2014). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing.

Reilly, F. K. (1985). *Investment Analysis and Portfolio Management* (Second ed.). San Francisco: Dryden Press.

Revelle, W. (2014). *psych: Procedures for Psychological, Psychometric, and Personality Research.* Evanston, Illinois: Northwestern University. R package version 1.4.5.

Rudin, A. and J. S. Morgan (2006). A portfolio diversification. *The Journal of Portfolio Management 32*, 81–89.

Statman, M. (1987). How many stocks make a diversified portfolio? *Journal of Financial and Quantitative Analysis 22*, 353–363.

Statman, M. (2004). The diversification puzzle. *Financial Analysts Journal 60*, 44–53.

Stevenson, R. A. and E. H. Jennings (1984). *Fundamentals of Investments* (Third ed.). San Francisco: West Publ. Co.

The Economist (2014). Will Invest for Food. [online] `http://www.economist.com/news/briefing/21601500-books-and-music-investment-industry-being-squeezed-will-invest-food`.

Wuertz, D., T. Setz, Y. Chalabi, and Rmetrics core team members (2013). *fBasics: Rmetrics - Markets and Basic Statistics*. R package version 3010.86.

Wuertz, D., T. Setz, Y. Chalabi, and R. C. Team (2014). *fPortfolio: Rmetrics - Portfolio Selection and Optimization*. R package version 3011.81.

Zheng, Z., B. Podobnik, L. Feng, and B. Li (2012). Changes in cross-correlations as an indicator for systemic risk. *Scientific Reports 2*, 888.

A Additional Tables

Table 4: The number of stocks in the selection pool in each two year sub-period.

Study Period	No. of stocks
2000-2002	171
2002-2004	172
2004-2006	175
2006-2008	187
2008-2010	195
2010-2012	190
2012-2014	194

Table 5: The KMO measure of sampling adequacy statistic for each two year sub-period.

	1st period	2nd period
2000-2002	0.51	0.54
2002-2004	0.58	0.50
2004-2006	0.45	0.65
2006-2008	0.75	0.86
2008-2010	0.77	0.73
2010-2012	0.81	0.90
2012-2014	0.61	0.71

Table 6: The 32 stocks that were selected from the 156 stocks used for the whole study period, based on a deletion criteria of an eigenvalue 1 and stop criteria of 0.64. The stocks that were retained in the 15 stock portfolio are highlighted.

Stocks Code	**Industry**
AGG	Basic Materials
MAH	Basic Materials
MDL	Basic Materials
RSG	Basic Materials
TRY	Basic Materials
AVG	Consumer Goods
ELD	Consumer Goods
GUD	Consumer Goods
MTS	Consumer Services
PRT	Consumer Services
SWM	Consumer Services
VRL	Consumer Services
AOG	Financials
BOQ	Financials
CPA	Financials
DJW	Financials
IBC	Financials
IOF	Financials
REA	Financials
RHC	Health Care
RMD	Health Care
AJL	Industrials
HIL	Industrials
MRM	Industrials
PMP	Industrials
SKE	Industrials
SLX	Industrials
AUT	Oil & Gas
MLB	Technology
SMX	Technology
HTA	Telecommunications
AGK	Utilities